

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Simon Struna

**Učinkovitost spletnega oglaševanja pri
uporabi vedenjskega ciljanja
uporabnikov**

DIPLOMSKO DELO
UNIVERZITETNI ŠTUDIJSKI PROGRAM RAČUNALNIŠTVO
IN INFORMATIKA

MENTOR: izr. prof. dr. Zoran Bosnić

Ljubljana, 2014

Rezultati diplomskega dela so intelektualna lastnina avtorja. Za objavlanje ali izkoriščanje rezultatov diplomskega dela je potrebno pisno soglasje avtorja, Fakultete za računalništvo in informatiko ter mentorja.

Besedilo je oblikovano z urejevalnikom besedil \LaTeX .

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo: Učinkovitost spletnega oglaševanja pri uporabi vedenjskega ciljanja uporabnikov.

Tematika naloge:

V diplomski nalogi naj kandidat analizira podatke, pridobljene od oglaševalske mreže, in ugotovi, ali z uporabo vedenjskega ciljanja pri spletnem oglaševanju lahko dosežemo boljšo učinkovitost pri spletnem oglaševanju. Vedenjsko ciljanje naj simulira s profiliranjem uporabnikov glede na njihove kratkoročne ali dolgoročne podobnosti. Uspešnost naj ovrednoti z merami za ocenjevanje učenja in svoj pristop naj primerja z izsledki iz sorodne literature.

IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisani Simon Struna, sem avtor diplomskega dela z naslovom:

Učinkovitost spletnega oglaševanja pri uporabi vedenjskega ciljanja uporabnikov

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvomizr. prof. dr. Zorana Bosnića,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 22. oktobra 2014

Podpis avtorja:

Za strokovno vodenje in vse nasvete se zahvaljujem mentorju izr. prof. dr. Zoranu Bosniću. Posebna zahvala pripada Domnu Koširju za motivacijo, potrpežljivost, strokovno pomoč in napotke pri izdelavi diplomskega dela. Zahvaljujem se podjetju iPROM, zlasti Tomažu Tomšiču za strokovno in tehnično pomoč pri pripravi anonimiziranih podatkov. Zahvaljujem se Tjaši, ki je lepo skrbela zame in me spodbujala tekom izdelave diplomske naloge. Hvala družini za vso podporo in motivacijo pri študiju, predvsem v napornih izpitnih obdobjih. Hvala tudi vsem mojim prijateljem in študijskim kolegom, ki so poskrbeli za sprostitev in lepe trenutke tekom študija.

Kazalo

Povzetek

Abstract

1	Uvod	1
1.1	Vedenjsko oglaševanje	2
1.2	Ciljano oglaševanje	3
1.3	Oglaševalske mreže	4
2	Pregled področja	7
2.1	Načini ciljanja	9
2.2	Podobne raziskave	12
2.3	Zasebnost uporabnikov	14
3	Opis podatkov	17
3.1	Oglaševalsko podjetje	17
3.2	Strežniški dnevniki	18
3.3	Priprava podatkov	20
3.4	Testni množici	22
4	Metoda za merjenje učinkovitosti	25
4.1	Opis metode	25
4.2	Uporabniški profili	26
4.3	Gručenje uporabnikov	28
4.4	Evalvacija	29
4.5	Uporabljene tehnologije	34

KAZALO

5	Rezultati	35
5.1	Podobnosti med uporabniki in oglasi	35
5.2	Izboljšanje učinkovitosti oglasov	36
5.3	Natančnost, priklic in F-mera	38
5.4	Primerjava s sorodnimi raziskavami	40
6	Zaključek	43
6.1	Nadaljnje delo	44
A	Shema podatkovne baze	51

Povzetek

V našem diplomskem delu smo želeli na podlagi predpostavk vedenjskega ciljanja in sorodnih raziskav preveriti, ali vedenjsko ciljanje res lahko izboljša učinkovitost spletnega oglaševanja. Za namene raziskave smo pridobili anonimizirane podatke iz slovenske oglaševalske mreže, nad katerimi smo izvedli naš eksperiment. Vsakega uporabnika smo na podlagi obiskanih strani in kliknjenih oglasov predstavili v obliki uporabniškega vektorja. Vzporedno smo izvajali dva neodvisna eksperimenta, saj smo želeli s primerjavo dobljenih rezultatov preveriti, ali se pri gradnji uporabniških profilov bolje obnesejo dolgoročni ali kratkoročni podatki. V nadaljevanju eksperimenta smo želeli potrditi osnovno predpostavko vedenjskega ciljanja, ki trdi, da imajo uporabniki, ki klikaajo enake oglase in obiskujejo enake strani, tudi podobne interese. Izvedli smo postopek gručenja, nato pa za posamezne oglase poiskali najprimernejše gruče ter simulirali prikazovanje oglasov zgolj tem gručam. Primerjali smo dvig učinkovitosti oglasov pred in po gručenju ter hkrati preverjali še karakteristike podobnosti, priklica in F-mero, s čimer smo želeli ugotoviti, kako učinkovita je lahko naša metoda gručenja. V našem eksperimentu smo pokazali, da lahko z uporabo vedenjskega ciljanja in uporabo kratkoročnih podatkov za zajem uporabnikovih interesov dvignemo učinkovitost spletnih oglasov za kar 746 %. Na koncu smo rezultate in ugotovitve primerjali še z izsledki iz sorodne literature. Dobljeni rezultati simulacij so bili v skladu s pričakovanji in so primerljivi, je pa vsem raziskavam skupno to, da so rezultati močno odvisni od količine podatkov in metod obdelave. V zaključku podamo še nekaj predlogov za izboljšavo naše metode in ideje za nadaljnje delo.

Ključne besede: oglaševanje, vedenjsko ciljanje, oglasi, gručenje.

Abstract

In our thesis we wanted to examine, based on assumptions of behavioral targeting and related research, whether behavioral targeting can really improve the effectiveness of online advertising. We conducted our experiment on anonymised information that we have obtained from a Slovenian advertising network. Based on previously visited web pages and clicked ads, we represented each user profile as a vector of interests. We carried out two independent experiments, as we wanted to examine whether long-term or short-term data is better for generating user profiles. Further, in our experiment we wanted to confirm the basic assumption of behavioral targeting, which claims that users who click on the same ads and visit the same sites also have similar interests. We implemented the clustering method and simulated displaying ads to different clusters in order to find the most appropriate cluster for the ad. We compared the increase in the effectiveness of advertisements before and after clustering and simultaneously checked the characteristics of similarities, recall and F-measure in order to find out how effective our method of clustering can be. We proved in our experiment that using behavioral targeting and short-term data for capturing user's interests, effectiveness of online advertisements can be raised as high as 746 %. Finally, we compared our results with the results of related literature. The obtained simulation results were in line with expectations and are comparable, but all research papers highlighted that the results are highly dependent on the amount of data and processing methods. In conclusion, we provided some suggestions to improve our methods and ideas for future work.

Keywords: advertising, behavioural targeting, ads, clustering.

Poglavje 1

Uvod

Z velikim naraščanjem uporabe interneta v zadnjih letih hitro naraščajo tudi interesi in razvoj spletnega oglaševanja. Internet ima zaradi velikih možnosti meritev in ogromnih količin razpoložljivih podatkov velik potencial kot oglaševalski medij. Razvoj in tehnologija na tem področju sta v zadnjih letih v naglem porastu, saj imajo spletni oglaševalci jasne in merljive cilje, vse razpoložljive informacije pa omogočajo bolj učinkovite možnosti za doseganje potencialnih kupcev. Naraščajo pa tudi želje po čim boljšem izkoristku vloženega oglaševalskega denarja in ena od možnosti za to je *vedenjsko ciljanje*.

Izraz *vedenjsko ciljanje* (angl. behavioural targeting) se nanaša na način spletnega oglaševanja, ki posameznikom prikazuje oglase na podlagi njihovega obnašanja in navad na internetu. Označuje skupek tehnologije in tehnik, ki jih uporabljajo spletni založniki in oglaševalci za izboljšanje odzivnosti njihovih spletnih oglaševalskih akcij. To dosega z zajemanjem in obdelavo podatkov, o tem kje in kako brskajo spletni uporabniki. Na ta način se lahko predvidi uporabnikova interesna področja, na podlagi katerih jim lahko prikazujejo bolj relevantne vsebine in oglase.

Poznamo več različnih vrst spletnega oglaševanja. V osnovi ga lahko delimo na dva glavna poslovna modela, na besedilne oglase v iskalnih omrežjih (angl. search network) in na oglase v prikaznem omrežju (angl. display network). Poznamo tudi oglaševanje po elektronski pošti (angl. email marketing), s katerim lahko dosežemo naslovnike na podlagi njihove vključitve v bazo elektronskih poštних naslovov. Z razvojem in naraščanjem priljubljenosti socialnih omrežij je v porastu

tudi oglaševanje v okviru teh portalov (angl. social media advertising), kot posledica napredka v mobilni telefoniji in razširjenosti njene uporabe pa se povečuje tudi investicija v oglaševanje na mobilnih napravah (angl. mobile advertising). Oglaševanje na spletu se tako razvija in prilagaja glede na njegovo uporabo in vedenje končnih uporabnikov. Raziskave kažejo, da posamezniki splet največ uporabljajo za elektronska sporočila, branje novic, časopisov ali revij ter iskanje različnih informacij [14].

Navkljub vedno večjim potrebam pa je področje vedenjskega oglaševanja še mlado in akademsko manj raziskano področje, sploh na področju oglaševanja v prikaznem omrežju s spletnimi pasicami (angl. banner advertising). To je bila naša motivacija, da v sodelovanju z oglaševalskim podjetjem izvedemo raziskavo in preverimo, ali vedenjsko ciljano oglaševanje res lahko izboljša učinkovitost spletnega oglaševanja. Za namene raziskave smo pridobili realne podatke iz slovenske oglaševalske mreže za obdobje enega tedna, na podlagi katerih bi lahko preverili in testirali različne metode ter prišli do konkretnih rezultatov o učinkovitosti vedenjskega ciljanja.

V našem diplomskem delu smo želeli na podlagi predpostavk vedenjskega ciljanja in sorodnih raziskav preveriti:

- Ali vedenjsko ciljanje res lahko izboljša učinkovitost spletnega oglaševanja?
- Ali imajo uporabniki, ki klikajo na podobne oglase, res tudi podobne interese in vedenjske vzorce na internetu?
- Ali je za napovedovanje uporabnikovega interesa in izgradnjo njegovega profila bolje uporabiti podatke krajšega ali daljšega časovnega obdobja?

1.1 Vedenjsko oglaševanje

Uporabniki z brskanjem po spletu, pregledovanjem vsebin in izvajanjem interakcij znotraj spletnih mest, za sabo puščajo sledi svojih aktivnosti. Spletni založniki in oglaševalci lahko nato uporabijo te podatke za izgradnjo uporabniških profilov in definiranje ciljne interesne populacije. Posamezne uporabnike združujejo v skupine z namenom prikazovanja prilagojenih vsebin glede na njihove interese. Ena od teh vsebin so seveda tudi oglasi. V teoriji naj bi pravilno ciljani oglasi

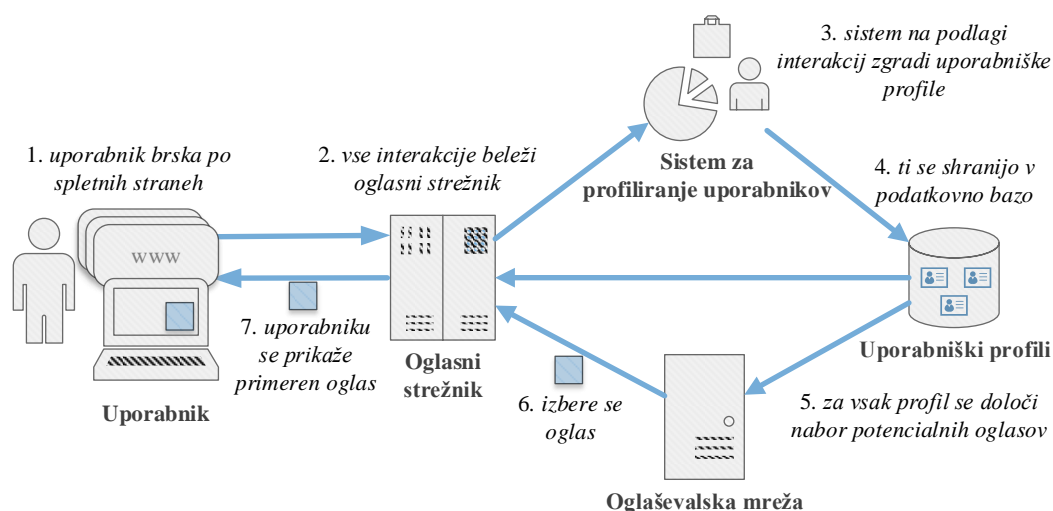
pri uporabnikih pritegnili več pozornosti, kar naj bi posledično vplivalo na boljše rezultate oglaševalskih akcij. Glavna ideja vedenjskega oglaševanja je ta, da bi uporabnikom kazali bolj primerne oglase s tematiko, ki jih bolj verjetno zanima. Na ta način naj bi na eni strani dosegli večji učinek oglasnega sporočila ter boljšo izrabo oglaševalskega proračuna oglaševalcev, na drugi pa manjšo vsiljivost oglasov za spletne uporabnike.

Za potrebe vedenjskega oglaševanja je potrebno spletne uporabnike prepoznati in jim na nek način slediti. Večina spletnih strani in oglaševalcev, ki uporablja vedenjsko ciljanje, zbira te podatke s pomočjo tekstovnih datotek, imenovanih HTTP piškotki (angl. cookies). Te se po uporabnikovi predhodni odobritvi zapišejo na uporabnikov računalnik. V tem piškotku je zapisana informacija, ki unikatno označuje določenega uporabnika. Vsi podatki o obiskanih straneh, klikih, iskanjih in dejanjih tega uporabnika na spletnih straneh se beležijo v oglasnem strežniku in se uporabijo za izgradnjo njegovega uporabniškega profila. Ko takšni uporabniki naslednjič obišejo spletno mesto znotraj spletne mreže, jim na podlagi teh podatkov lahko prikažejo oglas s primerno vsebino. Kadar se to izvaja brez vednosti uporabnikov, se obravnava kot kršenje zasebnosti uporabnika.

1.2 Ciljano oglaševanje

Ciljanje na osnovi vedenjskih vzorcev se izvaja s posebno programsko opremo - oglasnim strežnikom (angl. Ad server). Tehnologija za ciljanje na osnovi vedenjskih vzorcev pri prikazovanju oglasov na spletu upošteva preteklo aktivnost uporabnika na spletni strani. Če se glede na njegovo aktivnost ugotovi, da se uvršča med potencialne kupce določenega produkta, mu oglasni strežnik prikaže primeren oglas. S tem dosežemo, da je oglaševalec pokazal oglasno sporočilo potencialnemu kupcu, spletni uporabnik pa je videl oglasno sporočilo za izdelek, ki bi mu lahko ustrezal. Celoten postopek vedenjskega ciljanja je prikazan in opisan na sliki 1.1.

Postopek vedenjskega ciljanja se običajno deli na dva koraka. Prvi korak je izgradnja uporabniškega profila na podlagi njegovih preteklih dejanj (ogledane spletne strani, iskanja, kliki na oglase, čas ogledovanja itd.), na podlagi česar se ga umesti v določen segment oziroma skupino. V drugem koraku se iz nabora oglasov izbere tiste, ki so za določen segment najbolj primerni. Tako se za vsakega od



Slika 1.1: Prikaz postopka in opis posameznih korakov vedenjskega ciljanja

oglasov določi interesne skupine, za katere se predvideva, da so bolj zainteresirane za določene produkte. Posledično naj bi obstajala večja verjetnost, da uporabniki ciljanega segmenta kliknejo na takšen oglas in opravijo nakup ali naročilo storitve. Prikazovanje takšnih oglasov temelji na preteklem vedenju uporabnikov in že izkazanih interesih, kar naj bi vodilo k boljši učinkovitosti oglasnega sporočila.

1.3 Oglaševalske mreže

Spletna oglaševalska mreža je podjetje ali storitev, ki povezuje na eni strani oglaševalce in ponudnike oglasnega prostora (založnike) na drugi. Oglaševalske mreže uporabljajo vedenjsko ciljanje drugače kot posamezne spletne strani. Glede na to, da prikazujejo velik nabor oglasov preko velikega števila spletnih strani, imajo boljše možnosti gradnje spletnih profilov uporabnikov kot posamezne spletne strani. Na voljo imajo več podatkov, na podlagi katerih lahko sklepajo o demografskih lastnostih uporabnikov. Kot primer lahko vzamemo uporabnika, ki je obiskal spletne strani o nogometu, financah in spletni katalog za moško modo. Iz tega vzorca obiskanih strani lahko s precejšnjo verjetnostjo ugibamo, da je tak uporabnik moškega spola. Takšna analiza in gradnja uporabniških profilov omogoča,

da lahko oglaševalske mreže namesto oglaševanja na posameznih spletnih straneh oglaševalcem raje ponujajo oglaševanje točno določenim interesnim skupinam ciljne populacije (angl. audience targeting). Z uporabo te metode bi se na primer uporabniku, ki pogosto obiskuje turistična spletna mesta, prikazal oglas za počitniški paket, uporabniku, ki si večkrat ogleduje spletna mesta s pretežno športno tematiko, pa oglas za športni dogodek ali opremo.

Poglavje 2

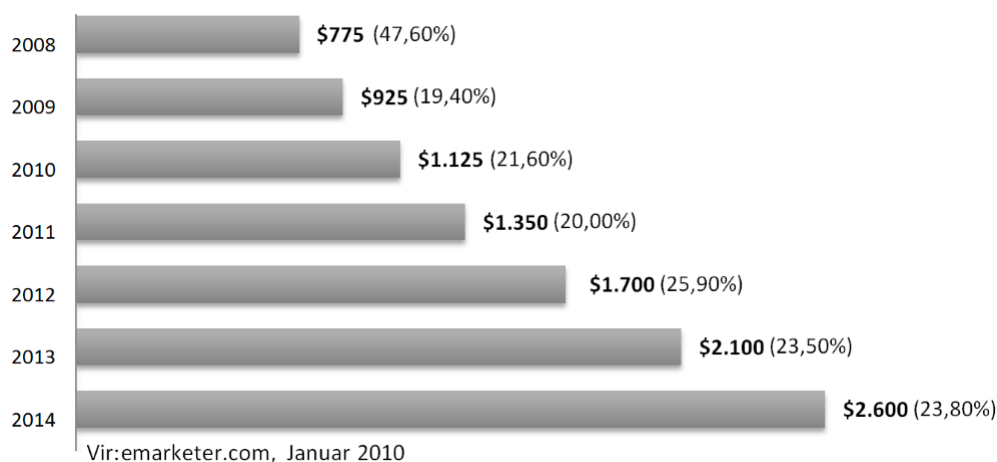
Pregled področja

Po nekaterih podatkih je bilo v letu 2013 na spletu skoraj 40 odstotkov svetovne populacije, najvišjo stopnjo razprostrtosti interneta, kar 75-odstotno, beležimo v evropskih državah [15]. Slovenija je glede na te podatke v evropskem povprečju, saj se je delež gospodinjstev z dostopom do interneta od leta 2007 do 2013 povečal s 57,61 odstotkov na 75,62 (545.284) [14]. S tem se povečuje tudi interes podjetij za oglaševanje na internetu, saj ima zaradi velikih možnosti meritev in ogromnih količin razpoložljivih informacij velik potencial kot oglaševalski medij.

Razvoj tehnologije na tem področju je v zadnjih letih v naglem porastu, saj internet kot oglaševalski kanal omogoča učinkovite možnosti za doseg potencialnih kupcev v izbrani ciljni skupini, spletni oglaševalci pa imajo na podlagi tega jasne in merljive cilje. V trenutnem obdobju ekonomske krize je poleg tega vse bolj prisotna potreba po čim boljšem izkoristku vloženega oglaševalskega denarja, pri tem pa je ena od možnosti za to vedenjsko ciljanje.

Oglaševalski trg v Evropi kaže znake rasti in inovacij na področju spletnega oglaševanja in marketinških investicij. Trendi kažejo, da se bodo v prihodnosti podjetja pri investiranju v spletno oglaševanje osredotočala predvsem na dobro poznavanje obstoječih in potencialnih strank ter njihovega vedenja [16].

Z razvojem obdelave podatkov in naprednejšimi metodami podatkovnega rudarjenja (angl. data mining), postaja vedenjsko ciljanje vedno bolj učinkovito in uporabljano. Od prvega pojava vedenjskega oglaševanja do danes je njegova uporaba strmo narasla. Na sliki 2.1 lahko vidimo naraščanje finančnih vložkov v oglaševanje z vedenjsko ciljanimi oglasi na ameriškem trgu od leta 2008. Za leto



Slika 2.1: Trend finančnih vložkov v vedenjsko ciljano oglaševanje v ZDA od leta 2008 do 2014. Na grafu so predstavljeni letni zneski v milijon \$ ter odstotek letne rasti.

2014 spletna stran eMarketer napoveduje, da bo v ZDA skoraj vsak četrti oglas prikazan na podlagi vedenjskega ciljanja.

Glavni razlog za takšno rast je učinkovitost tako za oglaševalce kot založnike, saj nekatere študije [4, 3] kažejo, da je lahko vedenjsko ciljanje več kot dvakrat bolj učinkovito kot ostali načini ciljanja. V raziskavi organizacije NAI (networkadvertising.org), izvedene s strani Howarda Bealea, se je pri vedenjsko ciljanih oglasih v kupca spremenilo 6,8 odstotkov uporabnikov, ki so kliknili na oglas. Pri neciljanih oglasih je bil delež takšnih 2,8 odstotka. Posledično vedenjsko ciljani oglasi zaradi učinkovitosti pri oglaševalcih dosegajo tudi do trikrat višjo ceno [17]. Poleg tega je potrebno upoštevati, da se tehnologija še vedno razvija in postaja vedno bolj učinkovita.

V današnjem času podjetja svetovnega merila, kot so Amazon, Google in Facebook, uporabljajo vedenjsko ciljanje [18, 19, 20]. Medtem ko veliko podjetij uporablja tovrstne storitve le za prikazovanje primernejših vsebin svojim uporabnikom, pa podjetja kot so Google in Facebook ponujajo tudi vsem ostalim možnost oglaševanja na podlagi njihove tehnologije in podatkov. Z napredno tehnologijo, enostavnim uporabniškim vmesnikom in svojo cenovno politiko so približali spletno oglaševanje najrazličnejšim oglaševalcem in različnim oglaševalskim proračunom

[18, 19]. Na ta način je širši populaciji dosegljivo oglaševanje na njihovih platformah s spletnimi oglasi, ki se prikazujejo želeni interesni skupini na podlagi njihovega preteklega vedenja in interesov.

2.1 Načini ciljanja

Učinkovito spletno oglaševanje je odvisno od treh glavnih faktorjev: od konteksta, v katerem se oglas pojavi, uporabnikov, katerim je prikazan, in od vsebine oglašnega sporočila. Za namene te raziskave se bomo osredotočili zgolj na uporabnike, katerim lahko prikazujemo znan nabor oglasov v okviru znanih spletnih strani. Sama grafična podoba oglasov, vsebina in kontekst, v katerem se pojavljajo, presegajo okvire te diplomske naloge. Za uspešno ciljanje je potrebno definirati funkcijo za ciljanje (angl. *targeting function*), ki bo na podlagi uporabnika in njegove zgodovine brskanja napovedala, kateri oglas bi ga utegnil zanimati. Taka funkcija za vsakega uporabnika napove, ali mu je smiselno prikazati določen oglas ali ne. Oglasni strežnik lahko nato na podlagi rezultata te funkcije postreže uporabniku zanj primeren oglas. V praksi poznamo več načinov ciljanja oglasov, od enostavnejših na predizbranih straneh do naprednejših, kjer lahko ciljamo zgolj specifične uporabnike oziroma ciljno populacijo [5].

Strani Oglaševanje na predhodno izbranih spletnih straneh je preprosta in popularna oblika ciljanja. Oglaševalec za svoje oglase določi nabor strani, na katerih naj se njegovi oglasi prikazujejo. Navadno gre za spletne strani, ki so tematsko povezane z oglaševanim produktom. Za obiskovalce teh spletnih strani se predvideva, da jih bodo oglaševalčevi oglasi zanimali in se bodo nanje uspešno odzvali. Na ta način oglaševalec, ki prodaja avtomobile, želi prikazovati oglase na spletnih straneh o avtomobilizmu, kdor prodaja ženske čevlje pa na straneh o ženski modi. Ta oblika ciljanja je popolnoma neodvisna od preteklega obnašanja uporabnika, izkorišča pa zgolj samo eno informacijo - spletno stran, ki si jo uporabnik trenutno ogleduje. Metoda temelji na predpostavki, da obstaja smiselna povezava med vsebino spletne strani in uporabnikovim interesom. Primerjamo ga lahko s klasičnim oglaševanjem v tiskanih medijih [12]. Tam na primer oglaševalci želijo svoj oglas v tistih tiskanih medijih ali revijah, za katere predvidevajo, da jih prebirajo poten-

cialni kupci njihovega izdelka ali storitve.

Kontekstualno ciljanje Pri tej metodi ciljanja se oglasi prikazujejo na spletnih straneh, ki so tematsko povezani z oglaševanim produktom. Sistem za kontekstualno ciljanje išče relevantne ključne besede po vsebini spletne strani, uporabniku pa nato postreže z oglasi, ki so vsebinsko povezani z besedilom spletne strani. Na ta način bi se na primer oglas za nov model tekaških copatov prikazoval na spletnih straneh povezanih s športom (tekom), tekaških forumih, novicah povezanih z atletiko in podobnih tematsko povezanih straneh s športom. Sama vsebina in kontekst spletne strani se navadno določi s predhodno uporabniško klasifikacijo ali pa gre za iskanje določenih besed in besednih zvez znotraj samega besedila spletne strani.

Naprednejša oblika te metode je **semantično ciljanje**, kjer se z naprednimi postopki tekstovnega rudarjenja in semantično analizo vsebine skuša *razumeti* sam kontekst spletne strani. V tem primeru ne gre le za iskanje posameznih ključnih besed v besedilu, ampak za prepoznavanje vsebine spletnih strani in prikazovanje tematsko najbolj primernih oglasov.

Demografske skupine in uporabniški segmenti Oglaševalska stroka ima že veliko izkušenj s prepoznavanjem svoje ciljne populacije in oglaševanjem posameznim uporabniškim segmentom [12]. Uporabnike združuje v različne skupine glede na njihove interese in demografske lastnosti, tem skupinam pa se potem prikazuje primerne oglase. Uporabnike deli običajno v različne segmente na podlagi njihovih demografskih, geografskih in psihografičnih lastnosti. Na ta način so na voljo različne oblike ciljanja glede na starost, spol, lokacijo (država/mesto), ... Posamezne lastnosti se ugotavljajo na podlagi vsebine strani, ki jih uporabniki obiskujejo, raznih anket in raziskav ter iz podatkov, ki jih uporabniki delijo s spletnim mestom ob registraciji [19, 18].

Ta način ciljanja ima dve glavni pomanjkljivosti. Oglaševalci lahko prikazujejo oglase le že obstoječim predhodno izbranim segmentom, poleg tega pa segmenti njihove ciljne populacije pogosto ne zajamejo dovolj natančno. Običajno je namreč o posameznih interesih in trenutnih potrebah posameznika težko sklepati zgolj iz njegove starosti, spola in lokacije. Navadno se ta oblika ciljanja kombinira z drugimi metodami z namenom natančnejšega zajema primernih uporabnikov za

določeno oglasno sporočilo [19].

Vedenjsko ciljanje uporabnikov Vedenjsko ciljanje za napoved obnašanja uporabnikov v prihodnosti uporablja podatke o njihovem obnašanju v preteklosti. Funkcijo za ciljanje naučimo iz zgodovine njihovih preteklih dejanj [6]. Tehnike vedenjskega ciljanja se razlikujejo med sabo glede na število parametrov, ki jih upoštevajo v napovedi, in v ciljih, za katere skušajo optimizirati uspešnost napovedi. Metodo vedenjskega ciljanja si lahko predstavljamo kot proces, v katerem se *naučimo* funkcijo za ciljanje iz podatkov o uporabnikih in oglasih. Na podlagi uporabniških vektorjev zgrajenih iz zgodovine uporabnikov se postavi napovedni model. Ta lahko upošteva zelo bogat nabor različnih lastnosti uporabnikov in se lahko nauči, da direktno optimizira uspešnost oglasov, katera se lahko ocenjuje na podlagi razmerja prikazov in klikov ali različnih oblik konverzijskih ciljev. Ogrodje modela vedenjskega ciljanja je navadno dovolj splošno, da lahko za napoved upošteva še cel nabor ostalih informacij. Na ta način lahko upošteva kontekst strani, ki si jo uporabnik trenutno ogleduje, pripadnost različnim uporabniškim segmentom in v zadnjem času tudi podatke iz različnih socialnih omrežij [13, 18].

Na podlagi analize preteklega vedenja uporabnikov lahko ciljamo precej bolj napredne segmente uporabnikov. Na ta način bi lahko na primer proizvajalec osvežilnih pijač svoj novi produkt, za katerega je ugotovil, da bi najbolj zanimal segment polnoletnih mladostnikov in mladih staršev, oglaševal zgolj tema dvema segmentoma uporabnikov. Poleg demografskih podatkov (starost od 18 do 30, ustrezen spol), bi kriterije dodatno omejili z vedenjskimi vzorci. V primeru mladostnikov bi oglase prikazovali zgolj uporabnikom, ki pogosto iščejo fraze povezane s šolo, veliko uporabljajo klepetalnike in socialna omrežja ter so že kdaj izkazali interes za tovrstne produkte. V primeru mladih staršev pa bi izbrali tiste, ki se pogosto zadržujejo na straneh, povezanih z novorojenčki, starševskih forumih in člankih o starševstvu. Na ta način bi z oglasnim sporočilom še bolj natančno zajeli želeno ciljno populacijo.

Priljubljena oblika je metoda **ponovnega ciljanja** (angl. remarketing ali re-targeting), pri kateri se prikazuje določene oglase zgolj uporabnikom, ki so v preteklosti že obiskali oglaševalčevo spletno stran ali si na tej strani že ogledovali določen produkt. Tem lahko prikazujemo oglase s prirejenim oglasnim sporočilom

za storitev ali produkt, za katerega so v preteklosti že izkazali interes [7].

Cilj vedenjskega ciljanja je precej bolj napredno ciljati uporabnike, kot pa oglaševati zgolj posameznim segmentom uporabnikov (starost, spol, ...). Z učenjem algoritmov na realnih podatkih ter prepoznavanjem navad in interesov posameznih uporabnikov lahko pridemo do veliko bolj personalizirane uporabniške izkušnje, od katere imajo lahko koristi tako uporabniki kot oglaševalci.

2.2 Podobne raziskave

Leta 2009 je bila izvedena prva empirična raziskava vedenjskega ciljanja na podlagi realnih podatkov oglasov iz resničnega sveta [4]. V raziskavi so testirali učinkovitost vedenjskega ciljanja na podlagi strežniških dnevnikov besedilnih oglasov komercialnega spletnega iskalnika.

Na razpolago so imeli podatke za obdobje enega tedna (1. - 7. junij 2008). Ko so surove podatke očistili podatkov o robotih (vsi “uporabniki” z več kot 100 kliki dnevno) in izluščili le oglase v angleškem jeziku, so jim ostali podatki za 6.426.633 unikatnih uporabnikov in 335.170 unikatnih oglasov. Dodatno so izločili oglase z manj kot 30 zabeleženimi kliki v obdobju enega tedna, saj na podlagi njih ni mogoče priti do zanesljivih rezultatov raziskave. Eksperiment so izvajali nad preostalimi 19.901 različnimi oglasi.

Pri izgradnji uporabniških profilov so uporabili štiri različne strategije in podatke:

- dolgoročni podatki (obdobje 7 dni), na podlagi ogledanih strani;
- dolgoročni podatki (obdobje 7 dni), na podlagi iskalnih poizvedb;
- kratkoročni podatki (obdobje 1 dan), na podlagi ogledanih strani;
- kratkoročni podatki (obdobje 1 dan), na podlagi iskalnih poizvedb.

Izgradili so matriko, kjer vrstice predstavljajo uporabnike, stolpci posamezne spletne strani, posamezne vrednosti pa so izračunane na podlagi metode TF-IDF (podrobnosti te bomo spoznali v naslednjem poglavju). Podobne uporabniške vektorje so združevali v interesne skupine z metodo gručenja. Testirali so 4 različna števila gruč (20, 40, 80, 160) in dva algoritma gručenja (k-means in CLUTO). V tem koraku so raziskovali, v kolikšni meri lahko vedenjsko ciljanje resnično izboljša odzivnost spletnih oglasov (CTR). Medtem ko sta bila oba algoritma gručenja glede

rezultatov skoraj identična, so se kot precej bolj učinkoviti izkazali kratkoročni podatki (na podlagi iskalnih poizvedb) in delitev uporabnikov v 160 gruč. V tem primeru se lahko učinkovitost izboljša do 670 %, kar so potrdili tudi s Studentovim t-testom.

V raziskavi so prišli do treh pomembnih zaključkov: (1) Uporabniki, ki klikajo podobne oglase, imajo dejansko podobno vedenje na internetu; (2) Odziv oglasov (*CTR*) se, s pravilnim segmentiranjem uporabnikov v interesne skupine in vedenjsko ciljanimi oglasi lahko v povprečju izboljša tudi do 670 %; (3) Uporaba kratkoročnih podatkov brskanja za generiranje uporabniških profilov se izkaže za bolj učinkovite kot dolgoročni podatki.

Raziskavo je želel ponoviti Maciaszek [3] leta 2010, ki pa se je osredotočil zgolj na oglase iz prikaznega omrežja ("display network"). Večina akademskih raziskav na tem področju je namreč izvedenih na podatkih iz iskalnega omrežja - na besedilnih oglasih. Posebnost prikaznega omrežja je ta, da je zaradi grafične oblike oglasov (slika, Flash ali video) težje razbrati kontekst in vsebino kot pri tekstovnih oglasih. Klik na oglas ne razkriva takšne količine informacij o interesu uporabnika. V tem primeru se mora metoda vedenjskega ciljanja za izgradnjo interesnih profilov uporabnikov bolj zanašati na zgodovino in vsebino strani, ki so jih pregledovali. V raziskavi so se zgledovali po najbolj učinkovitih metodah iz [4].

Raziskavo so izvajali na podatkih enega dne (4. avgust 2010), vhodni podatki so pa bili precej bolj obsežni kot v primeru [4], saj so vsebovali 80 milijonov zapisov, 60.000 od teh so bili podatki o klikih. Informacije so bile zajete preko 140.000 različnih spletnih mest (domen) in 2.000 oglasov.

Testiranje so zaradi velike količine podatkov izvajali na porazdeljenem sistemu prek več strežnikov. Uporabili so model programiranja MapReduce na osnovi implementacije Apache Hadoop (<http://hadoop.apache.org/>). MapReduce poenostavlja paralelizacijo kalkulacij preko več računalnikov. Uporabljena podatkovna baza je bila Apache Hive (<http://hive.apache.org/>), ki je podobna SQL, le da je prilagojena za Apache Hadoop in omogoča porazdeljeni dostop do podatkov preko več računalnikov. Hadoop in Hive precej olajšata procesiranje velikih količin podatkov. Ker so potrebovali tovrstni sistem le za namene raziskave, so najeli fleksibilni sistem v oblaku podjetja Amazon, ki omogoča dinamično spreminjanje količine razpoložljivih sistemskih sredstev. Za primerjavo pohitritve avtor navaja, da bi

za vse izvedene kalkulacije na enem računalniku potrebovali 166 dni. Zahvaljujoč Amazonovemu oblaku so vse kalkulacije izvedli v parih tednih. Za izgradnjo uporabniških profilov so informacije združevali zgolj na nivoju domen. Uporabnike so razdelili na 160 segmentov (po zgledu [4]).

Na podlagi podatkov iz komercialnega oglasnega omrežja so prišli do zaključka, da lahko vedenjsko ciljanje izboljša odzivnost oglasov za več kot 900 %, kar so potrdili tudi s Studentovim t-testom.

2.3 Zasebnost uporabnikov

Ker vedenjsko ciljano oglaševanje za napoved interesov posameznih uporabnikov uporablja zbrane podatke o njihovem preteklem vedenju, je ta metoda dvignila veliko prahu v zvezi z zasebnostjo podatkov. Prav zato lahko vsi ponudniki tovrstnih storitev to počnejo le s predhodno odobritvijo uporabnikov.

Vsi ponudniki vedenjskega oglaševanja morajo poslovati v skladu z zakonodajo Evropske unije. V Sloveniji je bil 15. junija 2013 sprejet nov zakon o elektronskih komunikacijah ZEKom [9], ki ureja problematiko na tem področju. Ta je prinesel nova pravila glede shranjevanja informacij (piškotkov) in dostopa do informacij, shranjenih na računalniku ali mobilni napravi uporabnika. V skladu z Zakonom o elektronskih komunikacijah (*ZEKom-1*, 157. člen) je potrebno uporabnike za uporabo spletnih piškotkov obvestiti oziroma pridobiti njihovo privoljenje. Zakon se sklicuje na Zakon o varstvu osebnih podatkov (*ZVOP-1*), v skladu s katerim je dovoljeno zbiranje statističnih in arhivskih podatkov v anonimizirani obliki brez izrecne privolitve uporabnikov.

V Sloveniji smo sicer uvedli enega od strožjih oblik zakona o uporabi piškotkov, s čimer pa je možnost beleženja podatkov in sledenja uporabnikom v veliko primerih onemogočena. V veliko državah sveta pa temu ni tako, saj se lahko ponekod še vedno beleži uporabnike brez predhodnega dovoljenja [21]. Na ta način se lahko zabeleži veliko več informacij in na njihovi podlagi je tudi vedenjsko oglaševanje bolj učinkovito.

Sicer ne obstaja še veliko neodvisnih raziskav in trdnih dokazov o tem, koliko in na kakšen način lahko zbiranje teh informacij pomaga oglaševalcem in založnikom, so pa takšne raziskave pomembne pri upravičevanju tovrstnega početja

pri končnih uporabnikih. S prikazovanjem bolj primernih oglasov se namreč več njihova učinkovitost ter s tem vrednost takšnega oglasnega prostora [17]. To teoretično pomeni več denarja za založnike in posledično boljše vsebine za uporabnike. Na podlagi neodvisnih raziskav in rezultatov vedenjskega oglaševanja (med drugim tudi vsebine tega diplomskega dela), bi se založniki, oglaševalci, predvsem pa končni uporabniki lažje odločili, ali jim prednosti odtehtajo tovrsten poseg v zasebnost.

Uveljavlja se uporaba posebne ikone na vedenjsko ciljanih oglasih in gumb “Ne sledi” (angl. “Do not track”), s katerim bi lahko uporabniki imeli nadzor nad podatki, ki jih o njih zbirajo različne spletne storitve. Storitve naj bi omogočala pregled zbranih informacij ter hkrati blokado beleženja podatkov vsem ali zgolj posameznim oglasnim sistemom. Več informacij o tem in o primerih dobre prakse je dostopnih na spletni strani Youronlinechoices¹ ali Aboutads.info².

¹<http://www.youronlinechoices.com/goodpractice>

²<http://www.aboutads.info/>

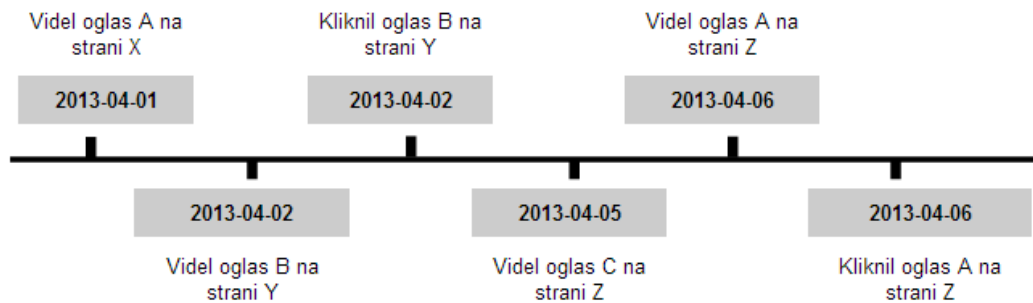
Poglavje 3

Opis podatkov

Za namene eksperimenta nam je oglaševalsko podjetje zagotovilo podatke njihove oglaševalske mreže. Podatke smo pridobili v obliki strežniških dnevnikov za obdobje od 8. 5. 2013 do 15. 5. 2013. Ti so bili zaradi varstva osebnih podatkov predhodno anonimizirani s postopkom, ki ni reverzibilen. Strežniški dnevniki vsebujejo podatke o zabeleženih uporabniških aktivnostih na spletnih straneh. Med njimi smo ločili dve glavni aktivnosti uporabnikov: podatke o prikazih in podatke o klikih na spletne oglase. Kadar je uporabnik na spletni strani izpostavljen oglasu v okviru oglaševalske mreže, se to v sistemu zabeleži kot prikaz ali klik oglasa. Zabeleženi podatki tvorijo časovno sled uporabnika o obiskanih straneh ter prikazanih in klikanih oglasih. Z obdelavo zapisov v strežniških dnevnikih smo te časovne sledi izluščili in jih zapisali v podatkovno bazo kot zaporedje dogodkov. Na ta način smo vsakega unikatnega uporabnika opisali kot zaporedje njegovih aktivnosti. Vsak zapis je vseboval datum in uro dogodka ter identifikator uporabnika, strani in oglasa. Pred začetkom eksperimenta smo podatke očistili zapisov o aktivnostih spletnih robotov in ostalih anomalij, ki bi lahko vplivale na rezultate. Pripravili smo dve ločeni vhodni množici podatkov, ki smo jih obravnavali v dveh neodvisnih eksperimentih. Na koncu smo primerjali rezultate obeh.

3.1 Oglaševalsko podjetje

Leta 1999 ustanovljeno podjetje je specializirano za načrtovanje in izvedbo spletnih oglaševalskih akcij ter je eno prvih podjetij na področju spletnega oglaševanja v



Slika 3.1: Primer zaporedja dogodkov v obliki časovne premice

Sloveniji. Njihove tehnološke rešitve so v skladu s sodobnimi standardi oglaševalske industrije [22].

Oglaševalsko mrežo za odkup neprodanih oglasnih kapacitet na spletnih straneh so predstavili leta 2000. Sistem, v katerega je aktivno vključenih preko 800 slovenskih spletnih mest, ponuja oglaševalcem različne oblike tehnološkega, kontekstualnega in vedenjskega ciljanja. Njihova tehnološka rešitev v oblaku omogoča centralizirano distribucijo, ciljanje in sledenje oglasnega gradiva na spletu.

3.2 Strežniški dnevniki

Vsak uporabnik pušča na spletu za sabo sled. Ta sled vsebuje podatke o tem, katere strani je obiskal, kaj je iskal ter kakšnim oglasom je bil izpostavljen in jih kliknil. Podatki o interakcijah so na spletnih strežnikih shranjeni v obliki strežniških dnevnikov (angl. server logs), ki predstavljajo najbolj osnovno obliko zapisa o spletnih aktivnostih. V teh datotekah je zapisan vsak zahtevek (angl. request) na oglasni strežnik, za vsakega od teh zahtevkov pa vsebujejo podatke o: uporabnikovem naslovu IP, datum/uro (angl. timestamp), zahtevani naslov, naslov vira (angl. referer), kodo HTTP, uporabnikov brskalnik in tehnične lastnosti njegove naprave. Iz tega nastane uporabnikova časovna sled (angl. time trail) v obliki zaporednih dogodkov (angl. events), ki opisujejo uporabnikovo obnašanje. Ponazorimo ga lahko s časovno premico kot na sliki 3.1. Vsaka vrstica strežniškega dnevnika predstavlja informacijo o enem takem dogodku, kateri je lahko:

- obisk strani (angl. pageview),

- ogled oglasa (angl. ad impression),
- klik na oglas (angl. ad click).

Za namene našega eksperimenta smo imeli na voljo anonimizirane strežniške dnevnike oglaševalske mreže iPROM Central za obdobje enega tedna. Namerno smo izbrali podatke iz obdobja pred uvedbo novega zakona o uporabi piškotkov, saj smo sklepali, da bodo bolj informativni od tistih, pridobljenih po uvedbi zakona. Problem od uvedbe zakona naprej so namreč *drugi* piškotki (angl. 3rd party cookies), saj ravno ti omogočajo sledenje enega in istega uporabnika preko več spletnih strani, na daljše časovno obdobje. Ob uporabi teh piškotkov je časovna sled uporabnika v strežniških dnevnikih precej daljša in posledično je o njegovih interesih in navadah na voljo več podatkov - več dogodkov v časovni sledi. Če uporabnik ne potrdi uporabe piškotkov, je ob vsakem obisku druge spletne strani zabeležen kot nov uporabnik.

Kot primer vzemimo uporabnika, ki se ni strinjal z uporabo piškotkov in je obiskal 10 spletnih strani v okviru oglaševalske mreže. V strežniških dnevnikih bi bilo v tem primeru namesto enega uporabnika, ki je obiskal 10 strani, zabeleženih 10 uporabnikov, vsak z obiskom le ene strani. To pa seveda ovira algoritme za napoved interesov uporabnikov. Idealno bi bilo, če bi lahko uporabnika sledili preko več strani, več brskalnikov in več naprav. Tako bi o njem imeli veliko več informacij in vedenjsko oglaševanje bi bilo lahko precej bolj natančno.

Strežniški dnevniki so vsebovali podatke o aktivnostih uporabnikov za obdobje od 8. 5. 2013 do 15. 5. 2013, njihova velikost je bila blizu 25 GB. Datoteke so skupno vsebovale $5,62 * 10^9$ vrstic.

Vhodne datoteke smo s skripto Perl najprej filtrirali, očistili raznih podatkov o robotih, odstranili uporabnike, ki so naredili več kot 100 klikov na oglase ter odstranili uporabnike, ki so se pojavili le enkrat - obiskali le eno stran ali videli le en oglas. Po tem postopku je naša vhodna množica vsebovala podatke za:

- 429.655 uporabnikov,
- 316 strani,
- 136 oglasov.

Zabeleženih smo imeli 31.971.449 prikazov in 10.629 klikov na oglase.

Po začetnih poskusih smo hitro ugotovili, da je taka količina podatkov prevelik zalogaj za naš testni sistem, saj so nekateri postopki prezahtevni. Nekateri vmesni izračuni parametrov namreč zahtevajo tudi do $3,41 * 10^{15}$ operacij, za izračun zgolj ene spremenljivke. Sledil je postopek obdelave strežniških dnevnikov in zapis pridobljenih podatkov v podatkovno bazo. Nadalje smo zmanjšali količino podatkov in pripravili dve vhodni množici podatkov, nad katerimi smo izvedli naš eksperiment.

3.3 Priprava podatkov

Za začetno obravnavo smo vzeli manjšo zalogo podatkov, s katerimi je bilo dovolj udobno delati in raziskovati, dela smo se lotevali postopoma. Računska zahtevnost nekaterih postopkov namreč hitro raste z večanjem količine podatkov. Korakoma smo povečevali količino podatkov in iskali meje našega postopka in testnega sistema. Ko smo ugotovili, koliko podatkov lahko obdelamo na našem sistemu, smo se lotili priprave dveh ločenih vhodnih množic.

S testiranjem in postopnim povečevanjem podatkov smo ugotovili, da lahko na našem testnem sistemu v dostojnem času obdelamo do 50.000 uporabnikov in 100 strani (ta dva parametra najbolj vplivata na zahtevnost izračunov, saj vplivata na velikost matrike uporabnik/stran). Podatke o oglasih smo v celoti ohranili, saj so to najbolj dragoceni podatki - sploh podatki o klikih. Te smo v celoti ohranili.

Ko smo razvili postopek, smo vse podatke iz strežniških dnevnikov vnesli v podatkovno bazo v tabelo `user_trail` - tabela 3.1.

Za namene testiranja smo želeli našo vhodno množico razdeliti na dva dela - na kratkoročne in dolgoročne podatke. Simulirati smo želeli gradnjo uporabniških profilov na podlagi podatkov za različna časovna obdobja. V prvi množici bi tako bili podatki za en dan, v drugi pa podatki za celoten teden. Na ta način smo naredili dve ločeni vhodni množici, ki smo jih obravnavali neodvisno in na njih ločeno opravili vse izračune. Na koncu eksperimenta smo primerjali rezultate obeh vhodnih množic. Na ta način smo ugotavljali, ali je pri gradnji uporabniških profilov bolj učinkovito zajeti uporabniške interese preko daljšega časovnega obdobja ali pa je dovolj upoštevati zgolj njihove najbolj ažurne preference.

Iz podatkov za cel teden smo izbrali naključen dan (12. 5. 2013) in podatke

Naziv	Primer vhodnega podatka
datum dogodka	2013-10-14
id uporabnika	1172736579
id strani	375
id oglasa	54687
ogledi oglasa	10
kliki oglasa	1

Tabela 3.1: Struktura obdelanih podatkov časovne sledi uporabnikov, pridobljenih iz strežniških dnevnikov.

	izvorno	teden	dan
zabeleženih prikazov	31.971.449	13.841.412	4.929.300
zabeleženih klikov	10.629	2.358	1.129
user_trail tabela	8.227.498	2.318.692	1.445.876
event_clicks tabela	9.115	2.053	971
event_impressions tabela	8.224.378	2.318.692	762.093
število uporabnikov	429.655	429.655	134.284
število strani	316	316	283
število oglasov	136	136	90

Tabela 3.2: Količina podatkov v izvorni in v obeh ločenih bazah pred postopkom priprave končnih vhodnih množic.

kopirali v dve ločeni podatkovni bazi. V eno bazo smo zapisali podatke za en dan, v drugo pa podatke za celoten teden. Količina podatkov pred postopkom redukcije je predstavljena v tabeli 3.2.

3.4 Testni množici

Pri našem eksperimentu smo bili omejeni predvsem z računsko in prostorsko zmogljivostjo našega testnega sistema (ta je predstavljen v poglavju 4.5). Količino podatkov, ki jo lahko našem sistemu obdelamo v doglednem času, smo ugotovili s testiranjem in postopnim povečevanjem testnega nabora podatkov. Prišli smo do ugotovitve, da lahko testni množici vsebujeta vsaka po 50.000 uporabnikov in 100 strani, število različnih oglasov pa bistveno ne vpliva na računsko zahtevnost. Sledil je postopek redukcije podatkov, potrebno je bilo zmanjšati količino obravnavanih uporabnikov in strani. Postopek smo začeli na zgoraj opisanih podatkih, izvajali smo ga ločeno. Najprej smo redukcijo izvedli na podatkih za en dan, nato še na množici s tedenskimi podatkih.

Postopek redukcije je bil sledeč:

1. Izberemo 100 najbolj obiskanih strani.
2. Odstranimo uporabnike z manj kot 10 ogledi strani in nobenim klikom,
3. Odstranimo uporabnike z manj kot 10 zabeleženimi prikazi oglasov in brez klikov.
4. Če omejitev obsega podatkov ni zadostna, izmed preostalih naključno izberemo 50.000 uporabnikov.

Postopek smo ponovili na podatkih za celoten teden, izbrali smo istih 50.000 uporabnikov in istih 100 strani. Po končanem postopku smo imeli dve različni vhodni množici, količina podatkov je prikazana v tabeli 3.3. Množici sta vsebovali enako število uporabnikov in enak nabor strani, razlikovali pa sta se v številu zabeleženih oglasov in oglasov s podatki o klikih. Slednji so zelo pomembni, saj dajejo najboljšo povratno informacijo o uporabnikovem interesu za posamezne oglase.

Z dvema ločenima vhodnima množicama, pripravljenima za nadaljnjo obdelavo, smo se lotili poskusa. Od tu naprej je potekal celoten eksperiment ločeno, za

	dan	teden
število uporabnikov	50.000	50.000
število strani	100	100
število oglasov	74	116
oglas s podatki o klikih	70	42

Tabela 3.3: Končna količina podatkov v obeh množicah

vsako množico posebej. Na koncu smo rezultate obeh primerjali in vrednotili.

Poglavje 4

Metoda za merjenje učinkovitosti

Pri pripravi naše metode smo se najprej lotili raziskovanja in iskanja podobnih že izvedenih raziskav. Tekom raziskovanja smo ugotovili, da ni še izvedenih veliko empiričnih raziskav na tem področju, sploh na področju prikaznega oglaševanja. V nadaljevanju sledi predstavitev osnovnih izhodišč in metod podobnih raziskav na tem področju, ki so predstavljale osnovo za razvoj naše metode. Sledi definicija naše metode eksperimenta in podroben opis posameznih korakov v tem postopku. Predstavljene so tudi posamezne evalvacijske metrike, s katerimi smo ocenjevali uspešnost naše metode in samo učinkovitost vedenjskega oglaševanja.

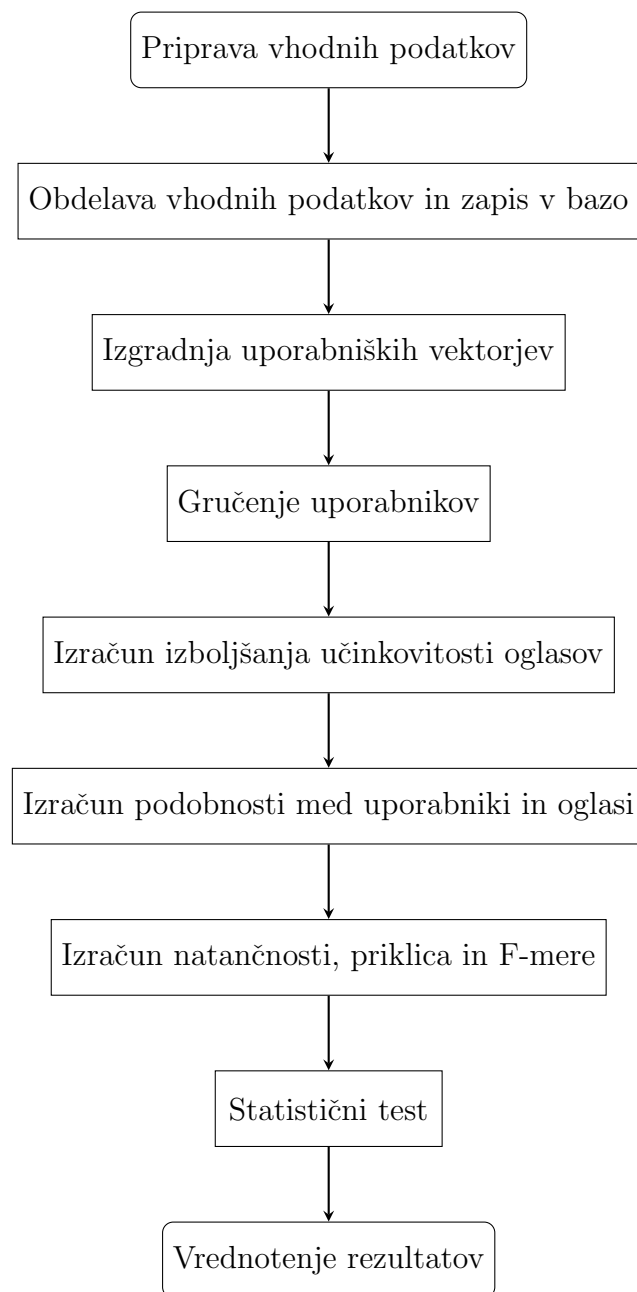
4.1 Opis metode

Pri pripravi podatkov in izvedbi raziskave smo se zgledovali po metodah podobnih raziskav iz poglavja 2.2. Pri načrtovanju eksperimenta smo se najprej lotili definiranja metode za obdelavo in pripravo vhodnih podatkov ter zapisa le teh v podatkovno bazo v primerni obliki. Iz surovih podatkov v obliki strežniških dnevnikov smo s postopki za njihovo obdelavo in metodo agregacije izluščili podatke primerne za zapis v podatkovno bazo. Pripravili smo dve ločeni vhodni množici podatkov, eno s tedenskimi podatki in eno s podatki za en dan. Od tu naprej smo vzporedno izvajali identičen postopek za vsako množico posebej. Sledil je

postopek izgradnje uporabniških profilov. Vsakega od uporabnikov smo skušali predstaviti v obliki vektorja, ki bi ponazarjal njegove interese na podlagi strani, ki jih je obiskal. Na podlagi teh smo se lahko lotili postopka gručenja. Cilj je bil odkriti podobne uporabnike s podobnimi intreresi in obnašanjem ter jih združiti v gruče. Po končanem postopku smo za vsakega od uporabniških vektorjev dobili informacijo, kateri gruči pripada. Lahko smo se lotili postopka simuliranja prikazovanja posameznih oglasov različnim gručam in izračuna izboljšanja CTR. Ker zgolj dokazati, da so lahko vedenjsko ciljani oglasi bolj učinkoviti ni dovolj, smo se po zgledu podobnih raziskav na tem področju lotili še dodatnih izračunov evalvacijskih metrik. S tem smo želeli dokazati in potrditi verodostojnost naših rezultatov. Sledil je postopek izračuna metrik podobnosti, natančnosti, priklica ter F-mere, opisanih v nadaljevanju. Po končanem postopku smo izvedli še statistični test, s čimer smo želeli potrditi rezultate našega eksperimenta. Celoten postopek v grobem prikazuje slika 4.1, podrobni opisi pa sledijo v nadaljevanju.

4.2 Uporabniški profili

Za predstavitev in opis obnašanja uporabnikov generiramo njihove profile glede na spletno strani, ki so jih obiskali. Uporabniki so predstavljeni v obliki vektorjev, kjer vsak vektor predstavlja drugega uporabnika, elementi posameznega vektorja pa predstavljajo posamezne spletne strani - njihovo obiskanost s strani uporabnika. Možnosti predstavitve obiskanosti spletnih strani v uporabniškem vektorju je kar nekaj. Med pogostejšimi omenimo frekvenčno porazdelitev, torej da vsako stran predstavimo kot število obiskov, diskretno vrednost (je obiskal/ni obiskal) ali pa normalizirano vrednost med 0 in 1 ($\#obiskov_strani_i / vsi_obiski$). V našem primeru smo uporabili metodo TF-IDF (angl. Term Frequency - Inverse Document Frequency), ki se običajno uporablja pri tekstovnem podatkovnem rudarjenju. Ta statistična vrednost se uporablja za opisovanje, kako pomembna je beseda znotraj množice dokumentov. Frekvenca besede (angl. TF - Term Frequency) označuje število pojavitev dane besede znotraj dokumenta, obratna frekvenca (angl. IDF Inverse Document Frequency) pa pove, kako redka je dana beseda v kontekstu vseh dokumentov. Obratna frekvenca daje nižjo utež najbolj pogostim besedam. V našem primeru je beseda stran, uporabnik pa dokument. Na ta način najbolj



Slika 4.1: Prikaz poteka eksperimenta

pogosto obiskane strani (naslovnice, iskalniki, ...) dobijo nižje uteži v vektorju, manj obiskane in bolj reprezentativne pa večje, saj naj bi bolje opisovale interes uporabnika.

Matematično gledano so vsi uporabniki opisani kot matrika realnih vrednosti $U \in R^{l \times g}$, kjer je g število vseh uporabnikov in l število vseh različnih strani v naših podatkih. Tako je vsak uporabnik u_i predstavljen kot vektor uteži TF-IDF:

$$u_{ij} = (\log(a) + 1) \times \log \frac{l}{b} \quad (4.1)$$

kjer je:

- $i=1,2,\dots, g$,
- $j=1,2,\dots, l$,
- a je število obiskov uporabnika i spletne strani j ,
- b je število različnih strani, ki jih je obiskal uporabnik i .

Oznake in simboli

Tukaj so predstavljeni in opisani matematični simboli, ki se uporabljajo tekom eksperimenta. Nabor n oglasov je predstavljen kot:

$$A = \{a_1, a_2, \dots, a_n\} \quad (4.2)$$

Naj bo $U_i = \{u_{i1}, u_{i2}, \dots, u_{im_i}\}$ množica tistih uporabnikov, ki so videli ali kliknili oglas a_i . Vhodni podatki (tabela 3.1) vsebujejo `user_id` in `ad_id`, ki določajo posameznega uporabnika in posamezni oglas. Uporabljajo se za določanje števila ogledov in klikov uporabnika za določen oglas.

Logična funkcija $\delta(u_{ij})$ definira klik uporabnika u_{ij} na oglas a_i :

$$\delta(u_{ij}) = \begin{cases} 1 & \text{če je uporabnik } u_{ij} \text{ kliknil oglas } a_i \\ 0 & \text{sicer} \end{cases} \quad (4.3)$$

4.3 Gručenje uporabnikov

Glavni cilj strategije vedenjskega ciljanje je združiti podobne uporabnike v skupine. Postopek gručenja nam bo v tem eksperimentu omogočal simuliranje prikazovanja oglasov različnim skupinam uporabnikov.

V našem eksperimentu smo za postopek gručenja uporabili algoritem k-means [11, 10]. Gre za standarden postopek, ki vsakega uporabnika uvrsti v eno samo gručo (za razliko od mehkega gručenja). Gre za razdelitev n uporabnikov v k gruče. Sam algoritem deluje tako, da najprej naključno izbere k središč in dodeli posamezne uporabnike najbližjemu središču, nato pa skozi N iteracij popravlja ta središča in jim dodeljuje uporabnike, dokler pripadnost posamezni gruči ne konvergira.

V našem eksperimentu smo tudi testirali kako različno število gruč (vrednost parametra k) vpliva na učinkovitost vedenjskega oglaševanja. Testirali smo različne vrednosti k ($k \in \{5, 10, 20\}$). Število iteracij N smo nastavili dovolj veliko, da so se izračuni stabilizirali, običajno je bilo to pri 500 ali več iteracij algoritma.

4.4 Evalvacija

Nekatere metrike, uporabljene v tem eksperimentu, so običajne v spletnem oglaševanju. Takšen je na primer CTR (angl. Click-through rate), ki pove, kako pogosto uporabniki, izpostavljeni oglasu, nanj tudi kliknejo. Je količnik med številom klikov in prikazov oglasa, izražen v odstotkih. Nas bo predvsem zanimalo njegovo povprečno izboljšanje. Dodatno bo ta raziskava vsebovala nekaj metrik, ki so bile uporabljene v podobnih eksperimentih in izhajajo s področja podatkovnega rudarjenja - to so mere podobnosti med uporabniki in oglasi, natančnost (angl. precision), priklic (angl. recall), razmerje R ter F-mera (angl. F-measure).

4.4.1 Podobnost med uporabniki in oglasi

Osnovna predpostavka vedenjskega tarčenja je ta, da imajo uporabniki, ki brskajo po podobnih straneh in iščejo podobne stvari, tudi podobne interese. Torej v tem primeru obstaja večja verjetnost, da bodo kliknili nek oglas, kakor pa uporabniki z drugačnim spletnim obnašanjem.

Naša prva meritev skuša utemeljiti predpostavko, da lahko vedenjsko ciljanje pomaga spletnemu oglaševanju. Podobnost med dvema uporabnikoma u_{ij} in u_{st} definiramo kot $Sim(u_{ij}, u_{st})$. Če je predpostavka vedenjskega ciljanja pravilna, potem mora biti ta podobnost med uporabniki, ki so klikali iste oglase, večja, kakor pa med uporabniki, ki so klikali različne oglase. V poglavju 4.2 smo že pred-

stavili uporabnike v numeričnem vektorskem prostoru. Torej lahko za računanje podobnosti med uporabniki uporabimo mero kosinusne podobnosti (angl. cosine similarity). Podobnost med dvema uporabnikoma je definirana kot

$$Sim(u_{ij}, u_{st}) = \frac{\langle u_{ij}, u_{st} \rangle}{\|u_{ij}\| \cdot \|u_{st}\|} \quad (4.4)$$

kjer \langle, \rangle označuje vektorski produkt in $\| \cdot \|$ vektorsko normo.

Z dobljeno kosinusno podobnostjo med uporabniki lahko definiramo dodatne mere podobnosti med oglasi. Znotraj-oglasno podobnost (angl. within-ad similarity) definiramo kot povprečno vsoto kosinusne podobnosti vseh uporabnikov, ki so kliknili dani oglas a_i . To nam bo pokazalo, kako so si podobni uporabniki, ki so kliknili isti oglas. Za vsak oglas a_i definiramo mero podobnosti $S_w(a_i)$ kot:

$$S_w(a_i) = \frac{2}{l_i(l_i - 1)} \sum_{\delta(u_{ij})=1} \sum_{\substack{\delta(u_{ij})=1 \\ t \neq j}} Sim(u_{ij}, u_{st}) \quad (4.5)$$

kjer je $l_i = \sum_j \delta(u_{ij})$ število uporabnikov, ki so kliknili oglas a_i .

Dodatna mera podobnosti, ki jo definiramo, je med-oglasna podobnost (angl. between-ads similarity), meri pa podobnost med uporabniki, ki so kliknili različne oglase. Da nam odgovor na vprašanje, kako podobni so si uporabniki, ki so kliknili različne oglase. Definirana je kot:

$$S_b(a_i, a_s) = \frac{1}{l_i l_s} \sum_{\delta(u_{ij})=1} \sum_{\delta(u_{st})=1} Sim(u_{ij}, u_{st}) \quad (4.6)$$

Kjer je:

- $l_i = \sum_j \delta(u_{ij})$ število uporabnikov, ki so kliknili oglas a_i ,
- $l_s = \sum_j \delta(u_{sj})$ število uporabnikov, ki so kliknili oglas a_s .

Na podlagi obeh zgornjih vrednosti definiramo še relacijo R med $S_w(a_i)$ in $S_b(a_i, a_s)$ kot:

$$R(a_i, a_s) = \frac{S_w(a_i) + S_w(a_s)}{2S_b(a_i, a_s) + d} \quad (4.7)$$

Relacija R se bo povečevala s povečevanjem znotraj-oglasne podobnosti S_w in zmanjševala z večjo med-oglasno podobnostjo S_b . Iz tega sledi, da večji kot bo R , bolj verjetno je naša metoda vedenjskega tarčenja dosegla želene rezultate. Možno je, da mera podobnosti S_b med nekaterimi oglasi zavzame vrednost 0. V

teh redkih primerih ni možno izračunati vrednosti R , saj ima mera takrat vrednost neskončno. Za take primere uporabimo dodaten parameter zelo majhne vrednosti ($d = 0,001$), ki zagotovi da je deljenje v zgornji enačbi vedno izvedljivo.

Na koncu uporabimo S_w , S_b in R da preverimo, kako podobni so si vsi oglasi med sabo znotraj naše testne vhodne množice. Najprej definiramo povprečje obeh mer podobnosti:

$$S_w = \sum_i \frac{S_w(a_i)}{n} \quad (4.8)$$

$$S_b = \sum_i \frac{S_b(a_i)}{n} \quad (4.9)$$

Iz njiju pa nato izračunamo končno povprečno razmerje R kot kvadratno povprečje vseh vrednosti R , med vsemi oglasi vhodne množice:

$$R = \sum_i \sum_s \frac{R(a_i, a_s)}{n^2} \quad (4.10)$$

4.4.2 Razmerje med prikazi in kliki

Razmerje med prikazi in kliki (angl. CTR - click through rate) je najpogostejši performančni indikator pri večini oglaševalskih kampanj. Visok CTR običajno označuje, da je uporabnik bolj zainteresiran za prikazani oglas in naj bi posvetil več pozornosti njegovemu oglasnemu sporočilu. Veliko oglaševalskih kampanj dandanes uporablja bolj sofisticirane indikatorje za merjenje uspešnosti, tako imenovane konverzije, ki označujejo, ali je uporabnik opravil željeno akcijo (opravi nakup produkta, registracija, prijava na novice, tiskanje kupona, ...). Smiselno bi seveda bilo ocenjevati uspešnost oglasov tudi po teh indikatorjih, kar pa presega okvire tega diplomskega dela.

V našem primeru smo CTR definirali kot število uporabnikov, ki so kliknili nek oglas, deljeno s številom uporabnikov, ki so ga zgoj videli oziroma mu bili izpostavljeni:

$$CTR(a_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta(u_{ij}) \quad (4.11)$$

Da lahko preverimo, če lahko gručenje uporabnikov res izboljša CTR oglasa, definiramo razmerje CTR oglasa a_i glede na posamezno gručo g_k kot:

$$CTR(a_i|g_k) = \frac{1}{|g_k(U_i)|} \sum_{U_{ij} \in g_k(U_i)} \delta(u_{ij}) \quad (4.12)$$

Kjer je $CTR(a_i|g_k)$ število uporabnikov v k -ti grupi $g_k(U_i)$. Po gručenju uporabnikov v k grup primerjamo CTR vsakega oglasa glede na vsako grupo da preverimo, če bi prikazovanje oglasa zgolj posamezni grupi izboljšalo njegov CTR.

Ko izračunamo CTR za vsak par oglas—grupa, lahko odgovorimo na vprašanje, koliko se lahko izboljša skupni CTR s ciljanjem oglasov na specifične grupe uporabnikov. To lahko dosežemo tako, da vzamemo maksimalni CTR posameznega oglasa preko vseh grup in izračunamo povprečno vsoto:

$$\Delta(a_i) = \frac{CTR(a_i|g_*(a_i)) - CTR(a_i)}{CTR(a_i)} \quad (4.13)$$

kjer je $g_*(a_i)$ grupa uporabnikov s najvišjim CTR za dani oglas a_i :

$$g_*(a_i) = \arg \max_{k=1,2,\dots,K} CTR(a_i|g_k(a_i)) \quad (4.14)$$

Tukaj je pomembno izpostaviti dejstvo, da ni nujno, da ima grupa z najvišjim CTR za dani oglas a_i tudi največ prikazov za ta oglas. Več o tem v poglavju 5.2.

4.4.3 Natančnost, priklic in F-mera

Obstaja več načinov, s katerimi lahko ocenimo zanesljivost izračuna izboljšanja CTR. Eden od običajnih načinov za preverjanje kvalitete zajetih podatkov je računanje natančnosti (angl. precision) in priklica (angl. recall). Če obravnavamo uporabnike, ki so kliknili oglas, kot pozitivne primere in uporabnike, ki so videli oglas, vendar niso kliknili nanj kot negativne primere, potem izračunamo natančnost kot:

$$Pre(a_i|g_k) = CTR(a_i|g_k) \quad (4.15)$$

Natančnost je v tem primeru enaka številu uporabnikov v dani grupi g_k , ki so kliknili na oglas, deljeno s številom vseh uporabnikov (v grupi g_k), ki so videli ta oglas - kar je v bistvu CTR oglasa v danem segmentu.

Priklic pove kolikšen odstotek vseh uporabnikov, ki so kliknili nek oglas, je znotraj dane grupe uporabnikov. Definiran je kot število uporabnikov iz grupe g_k , ki so kliknili oglas a_i , deljeno s številom vseh uporabnikov iz celotne vhodne množice, ki so kliknili na ta oglas:

$$Rec(a_i|g_k) = \frac{\sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})}{\sum_{j=1}^{m_i} \delta(u_{ij})} \quad (4.16)$$

kjer je m_i število uporabnikov v naši vhodni množici. Večji kot je priklic, večja je verjetnost, da zajamemo s postopkom segmentacije uporabnike, ki so kliknili oglas a_i .

Visoka vrednost natančnosti označuje, da so gruče tesno grupirane okoli uporabnikov, ki so kliknili dani oglas. Priklic nam pokaže, kolikšen delež vseh klikov na dani oglas je bilo vključenih v dano gručo.

Na primer, da bi funkcija za ciljanje za 30 uporabnikov napovedala klik na specifičen oglas, od katerih bi jih le 20 res kliknilo, za 40 preostalih, ki so tudi kliknili oglas, pa ne bi napovedala klika. V tem primeru je natančnost take funkcije $20/30 = 2/3$, medtem ko je njen priklic $20/60 = 1/3$.

Obe postavki smo nato združili v harmonično sredino natančnosti in priklica. Le-ta se imenuje F-mera (angl. F-measure):

$$F(a_i|g_k) = 2 \times \frac{Pre(a_i|g_k) \times Rec(a_i|g_k)}{Pre(a_i|g_k) + Rec(a_i|g_k)} \quad (4.17)$$

Pokaže nam, kako dobro se naše gručenje obnaša za dani oglas a_i in gručo g_k . Poznana je tudi kot F_1 mera, ker sta natančnost in priklic enakomerno uravnoteženi. Dve pogosto uporabljani izvedenki sta še F_2 mera, ki dvakrat bolj uteži priklic kot natančnost, in pa $F_{0.5}$, ki da več poudarka na natančnosti.

F-mere vseh gruč in oglasov lahko združimo v skupno F mero, povprečeno preko vseh oglasov in performančno najboljših gruč:

$$F = \sum_{i=1}^n \frac{F(a_i|g_*(a_i))}{n} \quad (4.18)$$

4.4.4 Statistični test

Na koncu eksperimenta smo izvedli statistični test za potrditev naših rezultatov. Statistično smo primerjali rezultate za en dan in teden ter s tem preverili, če je razlika med vhodnima množicama signifikantna. Najbolj primeren za to je Wilcoxonov predznačeni rangirni test (angl. Wilcoxon rank-sum test), poimenovan tudi Mann-Whitneyev test. Pri tej metodi primerjamo dva neodvisna vzorca. Uporabljamo ga za testiranje ničelne hipoteze o enakosti aritmetičnih sredin dveh populacij, ne

da bi predpostavljali da sta populaciji normalno porazdeljeni. Predpostavljamo pa, da sta populaciji zvezni. Nadomešča nam parni t-test, uporabljamo pa ga v primerih odvisnih vzorcev in kadar ni izpolnjen pogoj normalne porazdelitve.

Ničelna hipoteza je, da sta aritmetični sredini obeh populacij enaki, nasprotna hipoteza pa je, da sta različni. Wilcoxonov test nam pomaga ovrednotiti pridobljene rezultate, v programskem jeziku R ga zaženemo z ukazom:

```
wilcox.test(x, y, paired=FALSE);
```

4.5 Uporabljene tehnologije

Testno okolje vsebuje strežniški sistem z osem jedrnim procesorjem Intel Xeon E5-2687W @ 3.10GHz, s 16GB RAMa in 512GB SSD diska. Na njem teče strežniška verzija operacijskega sistema Debian Linux. Podatkovna baza je MySQL, struktura baze v prilogi 1. Obdelava strežniških dnevnikov in njihov zapis v bazo, priprava vhodnih podatkov za gručenje, obdelava dobljenih podatkov z gručenjem ter računanje kosinusne podobnosti poteka v programskem jeziku Perl. Za namen gručenja se uporablja programsko orodje WEKA, napisano v programskem jeziku Java. Statistični Wilcoxonov test smo izvedli v programu R.

Poglavje 5

Rezultati

V tem poglavju bomo predstavili rezultate našega eksperimenta in njihovo interpretacijo. V poglavju 5.1 bomo obravnavali podobnost med uporabniki in s tem skušali potrditi osnovno predpostavko vedenjskega oglaševanja. V poglavju 5.2 bomo predstavili faktor izboljšanja z vedenjsko ciljanimi oglasi, kar bomo skušali nadalje potrditi z rezultati v poglavju 5.3 in s statističnim testom. Na koncu bomo v poglavju 5.4 naredili povzetek pridobljenih rezultatov in jih skušali primerjati z rezultati podobnih raziskav.

5.1 Podobnosti med uporabniki in oglasi

Z računanjem podobnosti med uporabniki in oglasi smo želeli potrditi, da so si uporabniki, ki klikajo iste oglase, res bolj podobni kot uporabniki, ki klikajo različne oglase. S tem bi dobili potrditev, da imajo uporabniki s podobnim obnašanjem na internetu res tudi podobne interese. Ker se uporabnikovi interesi hitro spreminjajo, smo izvedli dve neodvisni testiranji, z dvema naboroma podatkov. Želeli smo namreč preveriti, kaj se bolje obnese pri gradnji uporabniških profilov - zgodovina obnašanja uporabnikov za krajše ali daljše obdobje. Za kratkoročne podatke smo tako obravnavali podatke enega dne, za dolgoročne pa podatke enega tedna.

V tabeli 5.1 vidimo rezultat meritev povprečnih vrednosti podobnosti preko vseh oglasov iz dveh ločenih naborov podatkov. Podobnost S_w označuje uporabnike, ki so klikali iste oglase, S_b pa podobnost med uporabniki, ki so klikali različne oglase. R označuje razmerje med njima.

	dan	teden
S_w	0,3017	0,1501
S_b	0,0425	0,0679
R	201,42	52,76

Tabela 5.1: Rezultati izračuna znotraj-oglasne (S_w) in med-oglasne podobnosti (S_b) ter razmerja med njima (R).

Kot vidimo, je vrednost S_w precej višja od vrednosti S_b v obeh primerih vhodnih podatkov. To potrjuje našo začetno hipotezo, da so si uporabniki, ki klikajo iste oglase, bolj podobni, kot pa uporabniki, ki klikajo različne oglase.

Visoka vrednost R pomeni, da je podobnost med uporabniki, ki klikajo iste oglase, več kot 200-krat večja kot med uporabniki, ki klikajo različne oglase. Opazimo še, da so boljši rezultat prispevali podatki za en dan, iz česar sklepamo, da je bolj smiselno graditi uporabniške vektorje na kratkoročnih podatkih.

Med našimi vhodnimi podatki je bilo več kot 95% oglasov z vrednostjo R večjo od 1. Lahko sklepamo, da velja zgornja trditev v večini primerov. Rezultate smo potrdili tudi s statističnim Wilcoxonovim testom ($W = 1256699937$, $p < 2.2e-16$).

Uspešno smo potrdili osnovno hipotezo, da so si uporabniki, ki klikajo iste oglase, bolj podobni kot uporabniki, ki klikajo različne oglase. Z vedenjskim ciljanjem torej lahko izboljšamo učinkovitost spletnega oglaševanja, nadalje pa smo želeli izmeriti dejansko izboljšanje, ki ga lahko dosežemo.

5.2 Izboljšanje učinkovitosti oglasov

Pomemben indikator povečanja učinkovitosti oglasov, predvsem gledano iz vidika oglaševalske mreže, je povečanje CTR oglasov. Kot smo opisali v poglavju 4, smo uporabnike najprej predstavili kot uporabniške vektorje na dva načina, na podatkih enega dne in enega tedna. Nato smo te uporabnike združevali v gruče z namenom testiranja, kako različno število gruč vpliva na dvig učinkovitosti oglasov. Uporabnike smo združevali v 5, 10 oziroma 20 gruč in primerjali CTR oglasov pred in po gručenju. Zanimalo nas je predvsem, kolikšen je odstotek izboljšanja, rezultati testiranja so predstavljeni v tabeli 5.2. V večini primerov se je izka-

	št. gruč	dan	teden
ΔCTR	$k = 5$	170,47%	162,61%
	$k = 10$	427,50%	375,95%
	$k = 20$	745,87%	681,87%

Tabela 5.2: Primerjava rezultatov izboljšanja CTR za različne parametre

zalo, da se kratkoročni podatki bolje obnesejo in dajo boljše rezultate. S tem, ko smo uporabnike razdelili v gruče, bi v primeru podatkov za en dan in $k=20$ gruč izboljšali CTR kar za 745,87%. Pri interpretaciji rezultatov pa moramo biti previdni, saj je pomembno upoštevati metode, po kateri so bili pridobljeni, in vhodne podatke, nad katerimi so bili opravljeni izračuni. Končno izboljšanje smo namreč izračunali kot povprečje najboljših vrednosti CTR oglasov v najboljših gručah. Vendar posamezni oglas lahko tako naredi zgolj majhen odstotek svojih prikazov v posamezni gruči, pa vendar hkrati v tej gruči dosega najboljši rezultat. Na primer, da ima oglas v eni od gruč 1 prikaz in 1 klik, je v tem primeru CTR oglasa za to gručo 100%, kar pa seveda ni smiselno upoštevati pri končnem izračunu. Ker so takšni rezultati lahko zavarajoči, smo filtrirali takšne nepravilnosti tako, da smo upoštevali rezultat v končnem izračunu izboljšanja CTR samo v primeru, da je oglas dosegel vsaj 50 prikazov oglasa v posamezni gruči.

V našem primeru so se za bolj učinkovite izkazali kratkoročni podatki, predvidoma zato, ker se danes uporabnikovi interesi in obnašanje na internetu hitro spreminjajo. Namreč, če skušamo uporabnikove interese zajeti preko daljšega obdobja, pri tem lahko izgubljam pomembno informacijo, kaj uporabnika v tem trenutku zanima. S kratkoročnimi podatki lahko najboljše zajamemo uporabnikove trenutne interese, razvoj metode za gradnjo kombiniranih uporabniških profilov, ki bi vsebovali tako kratkoročne kot dolgoročne podatke, pa bi bil ravno tako smiselno v okviru nadaljnjega dela.

Opazimo tudi, da se bolje obnese večje število gruč, v našem primeru je izboljšanje v primeru 20 gruč skoraj petkrat večje kot v primeru delitve uporabnikov na zgolj 5 gruč. Pomislili bi, da bi z večanjem števila gruč dosegli še boljše rezultate, vendar v našem primeru to ni bilo smiselno, saj smo z večjimi vrednostmi k dosegali slabše rezultate. V tem primeru gre torej bolj za iskanje najbolj optimal-

nega števila gruč, ki privedejo do najbolj učinkovitega izboljšanja za dani nabor podatkov. V našem primeru smo dobre podatke dobili pri 20 gručah.

Pokazali smo, da lahko s segmentacijo uporabnikov izboljšamo učinkovitost vedenjsko ciljanih oglasov. CTR oglasov se lahko dvigne do 745% s pravilno predstavitvijo uporabnikov in segmentacijo uporabnikov. Za nadaljno potrditev našega rezultata o izboljšavi CTR smo s statičnim testom primerjali CTR pred in po segmentaciji. Wilcoxonov test je pokazal, da so dobljeni rezultati statistično signifikantni ($W = 1934546166$, $p < 2,2e-16$).

5.3 Natančnost, priklic in F-mera

Učinkovitosti vedenjskega tarčenja ne moremo meriti zgolj z izboljšanjem razmerja med prikazi in kliki. Ni namreč dovolj le opaziti, da obstaja gruča uporabnikov, za katero velja

$$CTR(a_i|g_k) > CTR(a_i), \quad (5.1)$$

kjer $CTR(a_i|g_k)$ predstavlja število uporabnikov, ki so kliknili oglas a_i v gruči g_k . S tem zgolj dokažemo, da obstaja gruča uporabnikov, ki jih bolj zanima oglas a_i kakor ostale uporabnike, ne zagotavlja pa, da smo združili vse potencialne uporabnike v isto gručo.

Idealno bi bilo, da bi lahko vse uporabnike, ki so kliknili nek oglas, združili v eno gručo. V večini primerov pa žal ni tako, saj so zainteresirani uporabniki za določen oglas običajno porazdeljeni preko več gruč. V najslabšem primeru se lahko celo zgodi, da je vsak od uporabnikov v svoji gruči. Na ta način je težko izbrati najbolj primerno gručo za prikaz posameznega oglasa, metoda vedenjskega ciljanja pa v tem primeru ni ravno učinkovita. Iz tega razloga smo se želeli dodatno prepričati o natančnosti gručenja in možnostih optimizacije oglasov.

V poglavju 4.4.3 smo predstavili tri dodatne parametre, ki nam bodo pomagali bolje ovrednotiti naše izboljšanje CTR. Natančnost nam pove, kako dobro lahko zajamemo vse interesirane uporabnike za nek oglas znotraj ene gruče, priklic pa kako so takšni uporabniki, razpršeni preko več gruč. Oba parametra smo dodatno združili v harmonično vrsto in s tem dobili F-mero. Rezultate vsebuje tabela 5.3.

Natančnost nam pove, da je v povprečju slabih 5% uporabnikov takih, ki bi kliknili nek oglas znotra uporabniške gruče. Gre v bistvo za povprečni CTR znotraj

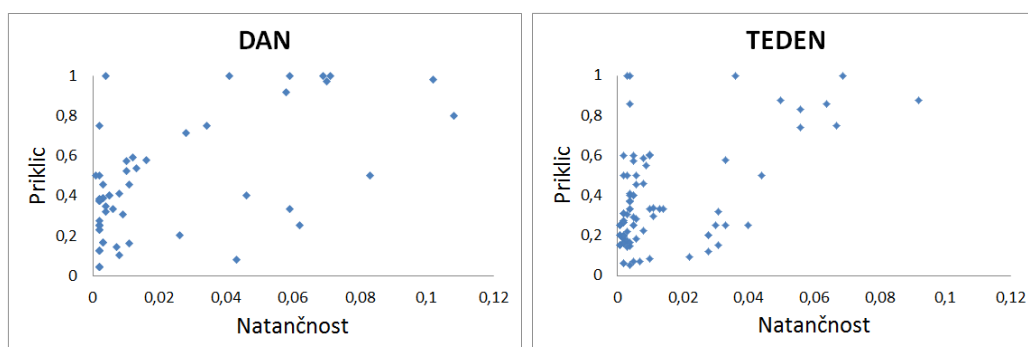
	št. gruč	dan	teden
Natančnost	$k = 5$	0,0156	0,0143
	$k = 10$	0,0276	0,0147
	$k = 20$	0,0451	0,0424
Priklic	$k = 5$	0,6512	0,60907
	$k = 10$	0,3887	0,4896
	$k = 20$	0,4089	0,36267
F-mera	$k = 5$	0,0292	0,02623
	$k = 10$	0,0488	0,0279
	$k = 20$	0,06150	0,05617

Tabela 5.3: Natančnost, priklic in F za različno število gruč

posamezne gruče. Sklepamo lahko, da so bili tem uporabnikom prikazani primerni oglasi. Po drugi strani ugotovimo, da pri 95% uporabnikov še obstaja možnost za izboljšavo in tem bi lahko pokazali bolj primerne oglase.

Z vrednostjo priklica dobimo informacijo o tem, kako podobni so si uporabniki znotraj ene gruče uporabnikov. V našem primeru je ta vrednost precej visoka (kar 0,65 v primeru podatkov za en dan in $k=5$). Iz tega sklepamo, da lahko v postopku gručenja v povprečju zajamemo veliko uporabnikov, ki bi kliknili nek oglas a_i . Torej bo več podobnih uporabnikov v istih gručah in bomo lahko dobro ciljali oglase. To visoko vrednost pripisujemo majhnemu številu uporabnikov, strani in oglasov v vhodni množici podatkov. Pri večji količini podatkov bi bila ta vrednost predvidoma precej manjša.

Na sliki 5.1 sta grafa primerjave natančnosti in priklica med vhodnimi podatki za en dan in teden. Vsak od grafov prikazuje vrednost natančnosti v odvisnosti od priklica, točke na grafu predstavljajo posamezne oglase. Že v prejšnjem poglavju smo pokazali, da so se za bolj učinkovite izkazali podatki enega dne, kar lahko vidimo tudi iz primerjave obeh grafov. Pri grafu za en dan imajo oglasi večje vrednosti natančnosti in nekoliko manjše vrednosti priklica. Iz tega sklepamo da je tudi izboljšanje CTR bolj učinkovito za ta nabor podatkov, saj lahko metoda gručenja bolj učinkovito zajame primerne uporabnike v posamezne gruče.



Slika 5.1: Primerjava vrednosti natančnosti in priklica za dan in teden. Posomezne pike na grafu predstavljajo oglase.

Mera F združuje zgornji dve vrednosti, pove pa nam, kako učinkovito je naše izboljšanje odziva CTR. Večja kot je vrednost, bolj z gotovostjo lahko sklepamo na učinkovitost našega izboljšanja z gručenjem. Tukaj velja poudariti, da se mera F lahko uporablja za vrednotenje učinkovitosti tudi v primerih, ko lahko en oglas prikazujemo različnim uporabniškim skupinam hkrati.

5.4 Primerjava s sorodnimi raziskavami

V našem diplomskem delu smo se zgledovali po metodah podobnih raziskav predstavljenih v poglavju 2.2. Da bi lahko čimbolje primerjali naše rezultate, smo se skozi naš eksperiment skušali držati postopkov, ki so bili predstavljeni v dokumentih sorodnih raziskav. To seveda zaradi same oblike vhodnih podatkov in načina zajema ni bilo mogoče v vseh korakih eksperimenta, smo pa lahko našli kar nekaj skupnih točk za primerjavo.

Glede na domeno našega problema in samo metodologije, so nam bili najbolj blizu rezultati raziskav [4] in [3]. Dočim so se v [4] ukvarjali tako z oglasi v prikaznem kot iskalnim omrežju, so se v [3] ukvarjali zgolj z analizo oglasov iz prikaznega omrežja, tako da nam je ta raziskava s tega vidika bolj sorodna. Po drugi strani so v [4], podobno kot v našem primeru, ugotavljali, kaj se bolje obnese pri izgradnji uporabniških vektorjev. Primerjali so podatke enega dne in enega tedna, kar je sorodno raziskavi, ki smo jo opravili mi. Primerjava rezultatov

			Yan[4]		Maciazsek[3]
	dan	teden	dan	teden	dan
Δ CTR	745%	681%	650%	606%	909%
Sw	0,3017	0,1501	0,2239	0,2594	0,2590
Sb	0,0425	0,0679	0,0196	0,0161	0,0410
R	201,42	52,77	44,29	91,19	1,40
Priklic	48,97%	38,87%	3,92%	8,54%	8,07%
Natančnost	2,76%	1,47%	9,09%	25,68%	83,40%
F	0,05	0,03	0,06	0,08	0,12

Tabela 5.4: Primerjava rezultatov s podobnimi raziskavami

je predstavljena v tabeli 5.4.

Podobnosti Podobnost med uporabniki, ki so kliknili iste oglase, je v našem primeru enega dne nekoliko višja od podobnih raziskav, za obdobje enega tedna pa nekoliko nižja. Višja je tudi vrednost S_b , kar pripisujemo manjšem številu oglasov. Dosegli smo visoko vrednost R in lahko trdimo, da je velika verjetnost da bo metoda gručenja posledično učinkovita.

Izboljšanje CTR Najboljši rezultate izboljšanja CTR smo v obsegu 909% dosegli v primeru [3] in je nekoliko višji kot v primeru [4]. Naše izboljšanje za teden je primelljivo s podatki za en dan v primeru [4], za izboljšanje za dan lahko rečemo, da je ravno v zlati sredini.

F-mera Opazimo, da so v primeru [3] imeli visoko stopnjo podobnosti in majhno vrednost priklica. Dosegli so precej višje vrednosti kot v primeru [4] ali v primerjavi z našim eksperimentom. Posledično so dosegli tudi precej višji F, lahko pa bi celo sklepali, da je posledica najvišjega F tudi najboljše izboljšanje CTR v primeru [3]. V našem primeru je vrednost F mere nekoliko nižja v primerjavi glede na vire [4] (0,15) in [3] (0,12), kar pripisujemo majhni količini vhodnih podatkov in manjši razpršenosti le teh. Čeprav je vrednost priklica visoka, nam nizka vrednost natančnosti zmanjša vrednost F. Vse 3 spremenljivke so pa tudi

precej odvisne od izbire parametra k v postopku gručenja.

Samo izboljšanje CTR oglasov je prav gotovo posledica več dejavnikov in potrebno je biti previden pri primerjavi teh rezultatov. Že sam način zajema in obdelave surovih podatkov, nato načina priprave vhodnih vektorjev, metode gručenja in nenazadnje tudi same karakteristike obravnavanega oglasnega sistema. Pogosto ne moremo govoriti o eni in edini pravi metodi, pogosto je namreč prava metoda skuppek nešteti eksperimentov in kombinacija metod, so pa takšne primerjave dober pokazatelj prave smeri raziskovanja in razvoja.

Poglavje 6

Zaključek

Cilj diplomske naloge je bil pokazati, da lahko z uporabo vedenjskega ciljanja pri spletnem oglaševanju dosežemo boljšo učinkovitost spletnih oglasov. Za ta namen smo razvili ustrezno metodo in pridobili realne podatke iz oglaševalske mreže, nad katerimi smo opravili eksperiment. Strežniški dnevniki za obdobje enega tedna so vsebovali zabeleženih 429.655 unikatnih uporabnikov, 316 strani in 136 oglasov - skupno 31 milijonov zabeleženih prikazov in več kot 10 tisoč klikov na oglase.

Vhodno množico smo najprej razdelili na dva dela, v enem so bili podatki enega tedna, v drugem za en dan. Nad njima smo vzporedno izvajali neodvisna eksperimenta, saj smo želeli s primerjavo dobljenih rezultatov preveriti, ali se pri gradnji uporabniških profilov bolje obnesejo dolgoročni ali kratkoročni podatki. V nadaljevanju eksperimenta smo preverjali, kako so si med sabo podobni uporabniki. S tem smo želeli potrditi osnovno predpostavko vedenjskega ciljanja, ki trdi, da imajo uporabniki, ki klikajo enake oglase in obiskujejo enake strani, tudi podobne interese. Izvedli smo postopek gručenja, nato pa za posamezne oglase poiskali najprimernejše gruče ter simulirali serviranja oglasov zgolj tem gručam. Primerjali smo učinkovitosti oglasov z dvigom CTR pred in po gručenju. Hkrati smo preverjali tudi karakteristike podobnosti, priklica in F-mero, s čimer smo želeli ugotoviti, kako učinkovita je lahko naša metoda gručenja.

Pokazali smo, da lahko z vedenjskim ciljanjem in ustreznim gručenjem uporabnikov izboljšamo učinkovitost spletnih oglasov za vsaj 746 %. Najboljši rezultate smo dosegli na podatkih enega dne, najbolj optimalno število gruč v našem primeru je bilo 20. Naše rezultate smo primerjali tudi z ugotovitvami podobnih raziskav in

prišli do zaključka, da so primerljivi, vendar je potrebno pri tem upoštevati obliko in metodo priprave vhodnih podatkov.

6.1 Nadaljnje delo

V našem diplomskem delu smo se lotili eksperimenta z uveljavljenimi in preizkušenimi metodami. Količino podatkov za obdelavo je omejeval naš testni sistem. V realnih sistemih, kot so oglasna omrežja, s povečano aktivnostjo na spletu količina informacij izredno hitro narašča. Vse te podatke je težko obdelati v realnem času, hkrati pa v njih prepoznati prave vzorce, ki nam bodo dali odgovore na naša vprašanja. Zelo natančno je potrebno definirati problem, za katerega iščemo rešitev, pravilno pripraviti vhodne podatke in razviti ustrezne metode, ki prostorsko in časovno ustrezajo naši problematiki. Pogosto razpoložljive informacije niso v primerni obliki ali pa ne vsebujejo pravih podatkov. Običajno tudi ni idealne metode, ki bi rešila vse probleme, ampak gre bolj za iskanje prave kombinacije ustreznih metod, filtriranje podatkov ter postavitve pravih indikatorjev, ki nam bodo dali odgovore na naša vprašanja. V nadaljevanju skušamo podati predloge za izboljšanje naše metode eksperimenta.

V našem diplomskem delu smo za gručenje uporabnikov uporabili preprosto metodo *k-means*. Predvidevamo, da bi z bolj naprednimi metodami gručenja lahko dosegli boljše rezultate. Uporabnikove interese smo v uporabniškem vektorju združevali le na nivoju domen. Smiselna bi bila gradnja uporabniških profilov na podlagi kategorij strani, katere bi klasificirali z avtomatično semantično analizo in kategorizacijo vsebine spletnih strani. Smiselno bi bilo tudi razviti adaptivno metodo, ki bi uporabnikove interese napovedovala iz ustrezne kombinacije njegovih kratkoročnih in dolgoročnih interesov. V nadaljevanju skušamo podati nekaj predlogov za nadaljnje delo.

Izgradnja uporabniških vektorjev Sklepamo, da so kratkoročni podatki bolj učinkoviti, ker se danes uporabnikovi interesi in obnašanje na internetu hitro spreminjajo. Če obravnavamo uporabnikovo obnašanje na daljši rok, se ga lahko na podlagi starih interesov umesti v napačne gručice in se mu posledično prikazuje vsebine, za katere že nekaj časa nima več interesa. Po drugi strani, kratkoročni

interesi izražajo uporabnikovo trenutno zanimanje in mu lahko pokažemo oglase, za katere je v tistem trenutku izkazal interes.

Smiselna bi bila gradnja kombiniranih uporabniških profilov, ki bi bili zgrajeni na podlagi dolgoročnih podatkov, stalno pa bi jih ažurirali z aktualnimi informacijami. Implementacija takšne metode v realnem sistemu bi namreč bila, da se na podlagi podatkov prejšnjega dne prikazuje primerne oglase naslednji dan. Uporabniški profil bi gradili postopoma (na podlagi tedenskih ali celo mesečnih podatkov) in ga prilagajali glede na spremembo interesov. V tem primeru bi bilo smiselno raziskati in ugotoviti pravo razmerje in metode za gradnjo ter posodabljanje takih profilov. Predvidevamo, da bi na ta način še dvignili učinkovitost vedenjsko ciljanih oglasov.

Pomembno je upoštevati dejstvo, da strani ki jih uporabnik obišče, niso vedno pravi pokazatelj njegovih interesov. Ko na seznamu zadetkov v iskalniku kliknemo enega od zadetkov, v večini primerov ne moremo vedeti, kam nas bo klik popeljal. V našem eksperimentu smo pri gradnji uporabniških profilov upoštevali zgolj število obiskov na posameznih straneh, ne pa tudi, koliko časa se uporabniki zadržijo na teh straneh, kar je zagotovo tudi pomemben podatek. Velikokrat vsebina za klikom namreč ni to, kar smo iskali, in v tem primeru takšno stran hitro zapustimo. Če bi beležili čas uporabnika na teh straneh (angl. *time on site*), bi lahko pri gradnji uporabniških profilov takšne strani izločili. Po drugi strani bi lahko bolj ovrednotili strani, na katerih se uporabniki dlje zadržujejo.

Gručenje CTR oglasov bi verjetno lahko še bolj izboljšali z uporabo bolj naprednih metod gručenja. V našem primeru smo uporabili preprosto metodo gručenja k-means, najboljše rezultate smo dosegli pri razdelitvi uporabnikov v 20 gruč. Dodatno povečanje števila gruč na našem obsegu podatkov ni bilo smiselno, bi se pa z večjo količino podatkov predvidoma lahko spremenilo tudi optimalno število gruč.

Prav tako v našem eksperimentu predvidevamo, da lahko uporabnik pripada zgolj eni gruči uporabnikov oziroma da je za nek oglas primerna le ena interesna skupina. V realni aplikaciji bi vpeljali ocenjevalno funkcijo, ki bi ocenila verjetnost pripadnosti uporabnika in primernost oglasa za neko skupino. Za vsak oglas bi pripravili nabor potencialnih uporabniških gruč (t.i. “mehko” gručenje) in v realnem času prilagajali prikazovanje glede na dejanske odzive.

Zajem uporabnikovih interesov Uporabnike smo v uporabniškem vektorju opisovali le na nivoju domen, ki so jih obiskali. Vemo pa, da se lahko vsebine znotraj enega spletnega mesta lahko močno razlikujejo. Tako bo nekdo na nivoju iste domene prebiral vsebine o športu, nekdo o politiki, spet nekdo drug pa kuharske recepte. Smiselno bi bilo vpeljati mehanizem kategorizacije vsebine. Tako bi za vsako domeno, rubriko, novico oziroma vsak URL naslov imeli informacijo, kakšno tematiko vsebuje njegova vsebina. Posledično bi lahko bolje sklepali na interes uporabnikov, ki te vsebine prebirajo. Postopka bi se lotili tako, da bi najprej določili glavne semantične kategorije. Te bi morale biti definirane tako, da hkrati zajamejo vse želje oglaševalcev po segmentirani populaciji, hkrati pa bi lahko vse spletne strani v našem sistemu umestili v vsaj eno od teh kategorij. Zgledovali bi se na primer lahko po taksonomiji, ki jo predlaga institucija IAB [23], ki poleg glavnih kategorij za opis vsebine vsebuje tudi veliko število podkategorij. Ko bi definirali kategorije, bi za vsak URL naslov določili ustrezno tematiko. To bi mogoče lahko storili ročno na nivoju domen, za vse vsebine znotraj spletnih strani pa je to preveč zahtevno. Da bi prepoznali vsebino posamezne spletne strani in jo avtomatično klasificirali z ustreznim mehanizmom, bi vsak naslov URL znotraj teh strani lahko obdelali z algoritmi in postopki za semantično analizo. Po tem postopku bi lahko uporabniške vektorje zgradili na nivoju kategorij. Predvidevamo, da bi takšen postopek prinesel boljše rezultate, saj bi bolje združil ustrezne uporabnike glede na njihove interese. V tem primeru bi v uporabniški matriki, predstavljeni v poglavju 4.2, vsaka vrstica še vedno predstavljala posameznega uporabnika, stolpci pa posamezne kategorije.

Po napovedih bo vedenjsko oglaševanje naraščalo z velikim zagonom in upamo, da bodo s takim tempom sledile tudi podobne akademske raziskave.

Literatura

- [1] I. Kononenko, *Strojno učenje*, Ljubljana: Založba fakultete za elektrotehniko in fakultete za računalništvo in informatiko, 2005.
- [2] I. Kononenko, M. Kukar, *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, Horwood Publishing, 2007.
- [3] R. Maciaszek, *How much behavioural targeting can help online display advertising?*. Birkbeck: University of London, 2010.
- [4] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, Z. Chen, “How much can Behavioural Targeting Help Online Advertising?”, v zborniku *Proceedings of the 18th international conference on World Wide Web*, Madrid, Španija, april 2009.
- [5] S. Pandey, M. Aly, A. Bagherjeiran, A. O. Hatch, P. Ciccolo, A. Ratnaparkhi in M. Zinkevich, “Learning to Target: What Works for Behavioral Targeting“ v zborniku *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, Glasgow, Škotska, 2011, str. 1805-1814.
- [6] Y. Chen, D. Pavlov in J.F. Canny, “Large-scale behavioral targeting“, v zborniku *KDD09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, Francija, 2009, str. 209-218.
- [7] A. Broder, V. Josifovski, *Introduction to computational advertising*. Dostopno na naslovu: <http://www.stanford.edu/class/msande239/>.
- [8] G. Hripcsak, A.S. Rothschild, “Agreement, the F-Measure and reliability in information retrieval”, v zborniku *Information Retrieval Journal of the American Medical Informatics Association*, 2005, str. 296-298.

- [9] (2014) Zakon o elektronskih komunikacijah (ZEKom), <http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO3781>.
- [10] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko in R. Silverman, "An efficient K-means clustering algorithm: Analysis and implementation", v zborniku *IEEE Transactions on Pattern Analysis and Machine Intelligence*, San Jose, ZDA, jul. 2002, str. 881-892.
- [11] JinHuaXu, HongLiu, "Web User Clustering Analysis based on KMeans Algorithm", v zborniku *2010 International Conference on Information, Networking and Automation (ICINA)*, Kunming, Kitajska, okt. 2010, str. V2-6 do V2-9.
- [12] Z. Jančič, V. Žabkar, *Oglaševanje*, Ljubljana: Založba FDV, 2013.
- [13] F. Provost, B. Dalessandro, R. Hook, X. Zhang in A. Murray, "Audience selection for on-line brand advertising: privacy-friendly social network targeting", v zborniku *Proceedings of the 15th SIGKDD international conference on Knowledge discovery and data mining*, Paris, Francija, 2009.
- [14] (2013) Statistični urad Republike Slovenije. *Uporaba interneta v gospodinjstvih in pri posameznikih, Slovenija, 2013 - končni podatki*. Dostopno na: http://www.stat.si/novica_prikazi.aspx?ID=5795.
- [15] (2013) International Telecommunication Union, *The world in 2013: ICT Facts and Figures*. Dostopno na: <http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx>
- [16] (2013) Forrester Research, *DigitalMedia Buying Forecast, 2012 to 2017*. Dostopno na: <http://www.forrester.com>
- [17] Howard Beales, *The value of behavioral targeting*, Advertising Institute Study, 2010. Dostopno na: <https://www.networkadvertising.org/>
- [18] (2014) Facebook.com, *How to target Facebook Ads*. Dostopno na: <https://www.facebook.com/business/a/online-sales/ad-targeting-details>

-
- [19] (2014) Google Ads, *Targeting Tools*. Dostopno na: <http://www.google.com/ads/displaynetwork/manage-your-ads/targeting-tools.html>
- [20] (2014) Amazon.com, *Interest-Based Ads*. Dostopno na: <http://www.amazon.com/b?ie=UTF8&node=5160028011>
- [21] (2014) Cookiepedia.co.uk. *Cookie laws accross Europe*. Dostopno na: <http://cookiepedia.co.uk/cookie-laws-across-europe>
- [22] (2014) iPROM d.o.o. *Optimizacija akcij na internetu*. Dostopno na: <http://www.iprom.si>
- [23] (2014) IAB Quality Assurance Guidelines (QAG) Taxonomy. Dostopno na: <http://www.iab.net/QAGInitiative/overview/taxonomy>

Dodatek A

Shema podatkovne baze

